molecules

MDPI

Article

Heterologous Machine Learning for the Identification of Antimicrobial Activity in Human-Targeted Drugs

Rodrigo A. Nava Lara [1], Longendri Aguilera-Mendoza [2], Carlos A. Brizuela [2], Antonio Peña [3] and Gabriel Del Rio [1,*]

[1] Department of biochemistry and structural biology, Instituto de Fisiología Celular, UNAM, Mexico City 04510, Mexico; rnava@email.ifc.unam.mx

[2] Computer Science Department, CICESE Research Center, Ensenada, Baja California 22860, Mexico; longendri@gmail.com (L.A.-M.); cbrizuel@cicese.mx (C.A.B.)

[3] Department of genetics, Instituto de Fisiología Celular, UNAM, Mexico City 04510, Mexico; apd@ifc.unam.mx

* Correspondence: gdelrio@ifc.unam.mx; Tel.: +52-55-5622-5663

Academic Editor: Julio Caballero

Received: 1 February 2019; Accepted: 14 March 2019; Published: 31 March 2019

Abstract: The emergence of microbes resistant to common antibiotics represent a current treat to human health. It has been recently recognized that non-antibiotic labeled drugs may promote antibiotic-resistance mechanisms in the human microbiome by presenting a secondary antibiotic activity; hence, the development of computer-assisted procedures to identify antibiotic activity in human-targeted compounds may assist in preventing the emergence of resistant microbes. In this regard, it is worth noting that while most antibiotics used to treat human infectious diseases are non-peptidic compounds, most known antimicrobials nowadays are peptides, therefore all computer-based models aimed to predict antimicrobials either use small datasets of non-peptidic compounds rendering predictions with poor reliability or they predict antimicrobial peptides that are not currently used in humans. Here we report a machine-learning-based approach trained to identify gut antimicrobial compounds; a unique aspect of our model is the use of heterologous training sets, in which peptide and non-peptide antimicrobial compounds were used to increase the size of the training data set. Our results show that combining peptide and non-peptide antimicrobial compounds rendered the best classification of gut antimicrobial compounds. Furthermore, this classification model was tested on the latest human-approved drugs expecting to identify antibiotics with broad-spectrum activity and our results show that the model rendered predictions consistent with current knowledge about broad-spectrum antibiotics. Therefore, heterologous machine learning rendered an efficient computational approach to classify antimicrobial compounds.

Keywords: machine-learning; antimicrobial peptide; non-peptidic antimicrobial compound; antimicrobial activity

1. Introduction

Drug-resistant microbes are one of the most important challenges for modern medicine [1] considering the increased rate in morbidity and mortality associated with antibiotic-resistant pathogens [2]. It is now commonly accepted that misuse of antibiotics is a major factor that promotes microbial resistance to these agents [3]; such is the case of broad-spectrum antibiotics that tend to promote resistance and are now prescribed in very restricted situations [4]. Furthermore, it has been noted that many non-antibiotic human-targeted drugs alter the gut microbiome in patients taking such drugs [5,6]. This alteration has been shown to be the consequence of a non-reported colateral antimicrobial activity, suggesting that microbe resistance to an antibiotic may emerge as a consequence of using those human-targeted drugs [7]. Furthermore, some antibiotics may have not been tested against the gut microbiome and may as well promote the emergence of resistant microbes. Since the experimental validation of antimicrobial activity for the gut microbiome requires tests on hundreds/thousands of cultivable and non-cultivable microorganisms and the number of new human-targeted drugs may include dozens of compounds, it is relevant to develop efficient computational strategies for the identification of secondary antimicrobial activity of human-targeted drugs. In the present work we present a computational strategy aimed to improve the identification of compounds with antimicrobial activity using machine-learning-based approaches.

Previous computational approaches to identify antibiotics using Quantitative Structure-Activity Relationships (QSAR) [8,9] and machine-learning-based [10,11] procedures have been reported. In these computational approaches, non-peptidic chemical compounds (from now on referred to as NPCC) are represented by chemical descriptors (e.g., *LogP*, molecular weight, polarizability) and each compound is labeled as antibiotic or non-antibiotic; then a clustering algorithm separates antibiotics from non-antibiotics. An important limitation of these previous studies is that the number of chemical compounds used to train the models is limited (less than one thousand NPCC have been described with antimicrobial activity) and the reliability of these models requires further improvement. Alternatively, antimicrobial peptides now accumulate in more than 10,000 in different databases [12–14], and several computational models have been reported to effectively classify antimicrobial from non-antimicrobial peptides [15–17]. Although peptides represent an important new focus to develop pharmaceuticals, most human-targeted drugs are NPCC; therefore computational models to identify antimicrobial activity in these compounds should focus on NPCC. The need to use common molecular descriptors between polypeptides and NPCC has been previously noted for protein-ligand recognition and protein folding, as a fundamental aspect to deal with induced-fit or conformer selection mechanisms for molecular recognition [18]; the aim of this work though, is not to find common descriptors to peptides and NPCC since there are already packages that solve this problem (see below). Here we propose that combining peptides and NPCC increases the training set size and this should improve the reliability of the computational models. The present work tests this proposal and validates the idea that heterelogous (NPCC and peptides) training sets render the best classifying models. We then show how this improved model may assist in the identification of broad-spectrum antibiotics on FDA-approved NPCC.

2. Results

2.1. Training and Testing Gut Antimicrobial Classifiers

Building data sets to combine peptides and NPCC required the use of molecular descriptors common to both types of compounds; in our case, we used 1444 descriptors calculated by PadelDescriptor (see Methods). Then, to identify the best machine-learning model to classify gut antimicrobials, three groups of training sets were used (see Table 1). The first group included only peptides (TrOnlyPeptides), the second group comprises 4 sets and included only NPCC (TrNPCC1-4) and the third group combined these two previous sets (TrHeterologous1-4) resulting in a total of 9 training sets (see Table 1); this rendered a total of 45 training sets. These 45 sets were further processed to substitute any null or "Infinity" values using three different approaches, and a reduction of dimensions was performed via principal-component analysis (PCA, see Methods). This procedure rendered a total of 50 Training Sets; all these sets are included in Supplemental Tables S1(A–E)–S9(A–E).

Nine testing sets were built using the NPCC recently reported by Maier et al. [7] with and without gut antimicrobial activity (see Table 2). The same processing of these testing sets was performed as in the case of the training sets (see above), rendering again a total of 50 data sets (see Supplemental Tables S10(A–E)–S18(A–E)). Please note that in both training and testing sets all peptides included were tested against only one gut microbe assayed against the NPCC used in these sets and that although there are many more peptides than NPCC in our training and testing sets, this imbalance is not relevant to find the border between antimicrobials and non-antimicrobials compounds.

Table 1. Training data sets.

Training Set	Entries	Description
TrOnlyPeptides	11,546	8000 antimicrobial peptides, 3546 peptides with no known antimicrobial activity
TrNPCC1	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity
TrNPCC2	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity
TrNPCC3	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity
TrNPCC4	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity
TrHeterologous1	6204	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2040 no antimicrobial compounds (1773 peptides and 267 non-peptidic compounds)
TrHeterologous2	6203	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2039 no antimicrobial compounds (1773 peptides and 266 non-peptidic compounds)
TrHeterologous3	6203	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2039 no antimicrobial compounds (1773 peptides and 266 non-peptidic compounds)
TrHeterologous4	6204	4164 antimicrobial compounds (4000 peptides and 164 non-peptidic compounds), 2040 no antimicrobial compounds (1773 peptides and 267 non-peptidic compounds)

The original NPCC from Maier et al. [7], here referred to as OnlyNonPeptides, was used to build TrNPCC1 by taking only the odd listed compounds, TrNPCC2 by taking even listed compounds, TrNPCC3 and TrNPCC4, included the first and second half of the data set respectively. The OnlyPeptides data set was divided to generate TrHeterologous1, TrHeterologous2, TrHeterologous3 and TrHeterologous4 by taking the odds listed peptides, even listed peptides, first and second half, respectively. Then, these TrHeterologous1-4 data sets with peptides were combined with the TrNPCC1-4 to complete these sets.

Table 2. Testing data sets.

Testing Set	Entries	Description
TeOnlyPeptides	861	328 antimicrobial and 533 non-antimicrobial non-peptides
TeNPCC1	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity. Same as TrNPCC2.
TeNPCC2	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity. Same as TrNPCC1.
TeNPCC3	431	164 antimicrobial non-peptides, 267 non-peptides with no known antimicrobial activity. Same as TrNPCC4.
TeNPCC4	430	164 antimicrobial non-peptides, 266 non-peptides with no known antimicrobial activity. Same as TrNPCC3.
TeHeterologous1	430	Same as TeNPCC1.
TeHeterologous2	431	Same as TeNPCC2.
TeHeterologous3	431	Same as TeNPCC3.
TeHeterologous4	430	Same as TeNPCC4.

The original NPCC from Maier et al. [7], here referred to as OnlyNonPeptides, was used to build all Testing Sets. TeOnlyPeptides was built taking all the 861 listed compounds. TeNPCC1 and TeHeterologous1 were built by taking only the even listed compounds. TeNPCC2 and TeHeterologous2 included only the odd listed compounds. TeNPCC3 and TeHeterologous3 included the second half of OnlyNonPeptides, TeNPCC4 and TeHeterologous4 included the first half of the data set. Testing sets were built so they were the complement of the compounds listed for their Training sets, so, for example, if a training set was built using the even listed compounds (e.g., TrNPCC1), its Testing set would be built with the odd listed compounds (e.g., TeNPCC1). Heterologous Testing Sets were the same as OnlyNonPeptides Testing sets, due to the fact that the interest compounds are of non-peptidic nature.

Five different statistical parameters (adjusted estimated error rate on the training set (AEER); correctly classified instances in the training set after splitting 33% for testing (%Split); 10-fold cross-validation (%10FCV); correctly classified instances on the testing set (%CC); area under the receiver operator characteristic curve on the testing set (AUROC)) that evaluated the performance on either the training or testing sets (see Methods) were used to identify the best classifier.

As shown in Figure 1, the best models included heterologous compounds (peptides and NPCC): circles in Figure 1 represent heterologous training sets and accumulate on the upper part of Figure 1, that is, those models with highest statistical parameters evaluating the model performance (the actual data in this figure for these models are included in Supplemental Table S19). Treating the training set rendering the best model with the K-nearest neighbor or mean-imputation approaches did not improve the performance of the best model (see Supplemental Tables S6G, S6I, S6J and its corresponding test set in Table S15G; supplemental Tables S6K, S6L and their corresponding test sets in Supplemental Tables S15I and S15J).

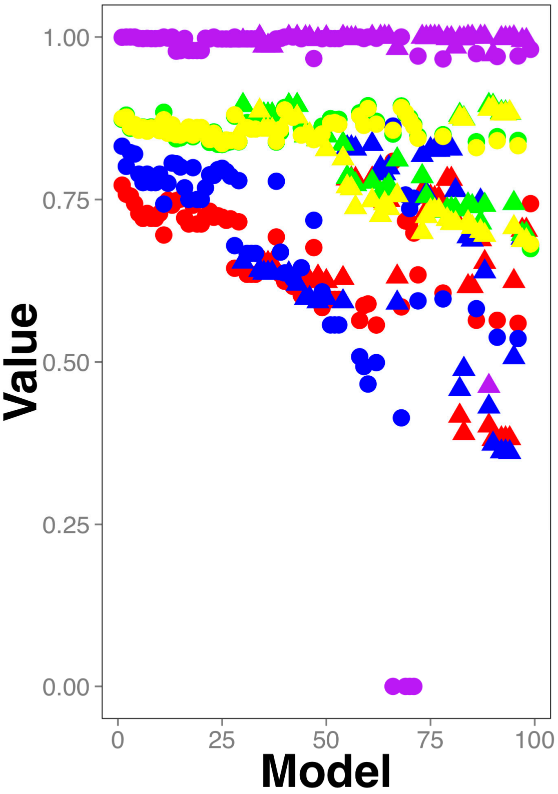

Figure 1. Classifiers performance. Five statistical parameters (yellow circle: Correctly classified instances in the test set after 67% of split validation of the training set; green circle: 10-fold cross-validation on training set; red circle: Correctly classified instances on the testing set; blue circle: AUROC on the testing set; purple circle: adjusted estimated error rate on the training set where the minimum error was represented by number 1.0) are sorted from highest to lowest values. Hence, the best parameter values are located on the left-upper part of the figure. The models using heterologous data are represented as circles; triangles are used otherwise. The actual data of this plot can be found in Supplementary Table S19.

Yet, none of these models surpassed the others in all 5 parameters. To aid in the visualization of this aspect of our results, Figure 1 displays the values in descending order from left to right; therefore, the models on the left side of the plot have better scores than those on the right. For instance, models using heterologous (represented by circles) testing sets (the red and blue circles, corresponding with the statistical parameters correctly classified instances and *AUROC*, respectively) laying on the left side of Figure 1, have better performance than those models using heterologous testing sets on the right side of the plot, yet, those on the right side including either heterologous or non-heterologous training sets (green and yellow circles or triangles) have better scores than those models using heterologous or non-heterologous training sets on the left side of the plot. The models in the middle of the plot have on the other hand, intermediate performances. Please note that the statistical parameter adjusted estimated error rate is the value that AutoWeka optimizes, hence for all the reported models is close to 1.0 and consequently does not contribute to differentiate the performance of the models. This statistical parameter is shown in Figure 1 to note that all models have similar error rates, yet different statistical parameters, hence, the best model obtained from AutoWeka cannot be selected simply by considering the error rate value reported.

Thus, to aid in the identification of the best models, we used a previous score developed by our group that takes into account multiple statistical parameters, the Combined Score or simply *CScore* [19]:

$$CScore_i = \frac{1}{5}\sum_{n=1}^{5}\left[\sqrt{\frac{MaxS_n - S_{i,n}}{MaxS_n - MinS_n}}\right] \tag{1}$$

where $MaxS_n$ and $MinS_n$ represent the maximum and minimum scores for a given statistical parameter n over all models; $S_{i,n}$ is the score observed for a given statistical parameter n and model i; n represents the index of the statistical parameter to evaluate (in our case were 5 parameters: AEER, %Split, %10FCV, %CC and AUROC). Thus, formula 1 calculates *CScore* for each model *i*.

CScore averages the difference of each statistical parameter to its best value (e.g., true-positive rate best value is 1, so the difference between the observed true positive rate and 1 is included in the *CScore*), therefore the lower the *CScore* value the better the classifying model. Figure 2 (and Supplementary Table S20) shows that the five best models are those using heterologous training sets (the ones below the 0.3 line in Figure 2). Furthermore, we noticed that the top 5 best models overlapped on average in more than 70% of their classifications hence, these were mainly redundant (see Figure 3).

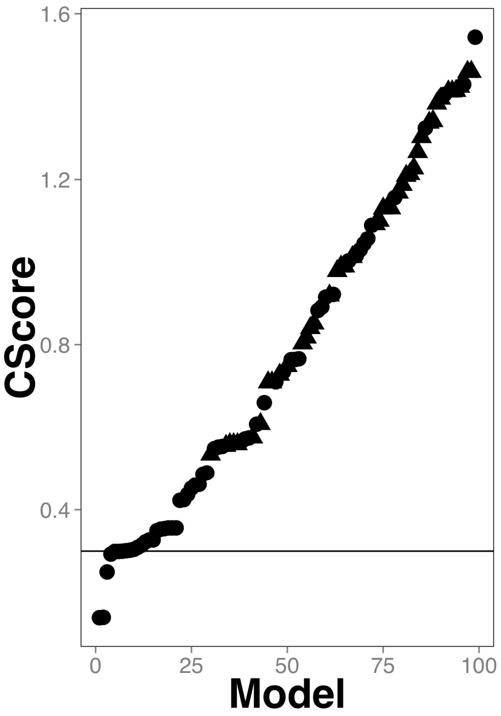

Figure 2. Classifiers combined scores. A circle represents each model; the best model has the lowest *CScore*. The line represents the *CScore* = 0.3, that separates the top 5 models from the rest. The models using heterologous data are represented as circles; triangles are used otherwise. The actual data of this plot can be found in Supplementary Table S20.

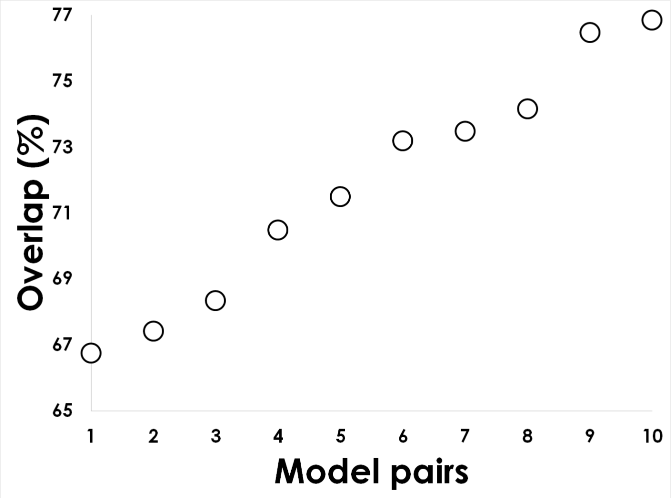

Figure 3. Classifiers overlap. The predictions of antimicrobial compounds of the top 5 models were compared to quantify their overlap. The image shows the 10 pairs of models generated from these 5 top models. The comparison was performed on the discovery set (see Methods) because not every model had the same testing set.

Therefore, we selected the best model based on the lowest *CScore*; such model was built using the RandomCommittee algorithm (see Supplemental Table S21 for the algorithm parameters) on the TrHeterologous1 set (see Table 1) that included 86 molecular PCA-reduced descriptors and achieved the following performance: AEER: 0.99955; %Split: 87.4; %10FCV: 87.3; %CC: 77.2; AUROC: 0.83 (model named TrHeterologous1-Reduced-With-99M-CRC20_CRF in Supplemental Table S21; the corresponding data set for this model is reported in Supplemental Table S6E).

2.2. Identifying Broad-Spectrum Antibiotics among FDA-Approved Compounds

We used the best model to predict NPCC with expected gut antimicrobial activity among FDA-approved drugs. The motivation to perform this prediction is not for testing purposes, as in the case of the training and testing sets used before. Hence, the set of compounds used in this prediction stage is referred to as the discovery set, because we aimed to discover potential compounds with gut antimicrobial activity. We used 756 FDA-approved compounds included in the ZINC database (see Methods) that were not part of the training or testing sets; these compounds included 111 antimicrobials and 645 compounds without any known antimicrobial activity; we also added 73 NPCC that included 22 antifungal compounds and 51 without any reported antifungal activity (see Supplementary Table S22). We have previously reported that these 22 antifungals work through a mechanism (alter calcium intake [20]) different from antibacterial compounds (e.g., penicillin derivates, sulphonamides, etc), thus we expected our model to predict few of these compounds as antibacterials. FDA-approved compounds on the other hand are expected not to have, or to have minor, gut antimicrobial activity otherwise their secondary gastrointestinal effects would be significant. We would expect that FDA-approved drugs would be less likely predicted to act against non-athogenic gut microbes than antifungals. To evaluate the reliability of our predictions using the discovery set, we considered that antibiotic compounds against the non-pathogenic gut flora among the FDA-approved drugs should be considered broad-spectrum antibiotics; please note that our classifier was not trained to predict this class of antibiotics, yet the combination of the predictions of our classifier on the FDA-approved drugs would render this information. The definition of broad-spectrum antibiotics is somehow arbitrary, for instance, it is considered that antibiotics that act on G(+) and G(−) are broad-spectrum antibiotics for some authors, while those acting against pathogenic and non-pathogenic microorganisms are classified as broad-spectrum antibiotics by others [21,22]. The list of broad-spectrum antibiotics was obtained from five recent works (see Methods), including 19 broad-spectrum and 3 narrow-spectrum antibiotics (see Supplementary Table S22). We were able to identify 72 true positives (FDA-approved antibiotics against pathogenic microbes predicted to act against non-pathogenic gut microbes) in the discovery set that we predicted should be considered as broad-spectrum antibiotics (see Table 3).

Table 3. Confusion matrix for the discovery set.

	Predicted Gut Antimicrobial	Predicted No Antimicrobial
Pathogenic antimicrobial	72	61
No antimicrobial	140	556

The actual data for this table can be found in Supplementary Table S22.

From these 72 antimicrobials, only 16 had been annotated as broad-spectrum antibiotics and 3 as narrow-spectrum antibiotics (see Supplemental Table S22). Hence, we propose that these 3 annotated narrow-spectrum antibiotics should be considered more likely as broad-spectrum antibiotics (see Table 4).

Table 4. True pathogenic antimicrobials predicted by the best classifier on the discovery set.

Compound Name	Annotation
Amoxicillin	Narrow spectrum
Phenoxymethylpenicillin	Narrow spectrum
Cephalexin	Narrow spectrum

On the other hand, among the 61 false negatives, 3 compounds were annotated as broad-spectrum antibiotics (see Supplemental Table S22). This annotation is consistent with our predictions, since these antibiotics directed towards pathogenic microorganisms are unlikely to affect the non-pathogenic gut microbes. Furthermore, 17 out of the 22 antifungal compounds were predicted as antimicrobials.

Thus, in total we were able to correctly identify 16 out of the 19 known broad-spectrum antibiotics and we suggest that 3 of the annotated narrow-spectrum antibiotics should be re-evaluated; hence, the reliability to identify broad-spectrum antibiotics was 84.2%. Furthermore, our results suggest that 56 (61 true negatives less 5 antifungals) (50.4%) out of 111 antibiotics approved by the FDA included in our discovery set are unlikely to affect gut microbes. In comparison, 5 (22.7%) out of 22 antifungals were predicted not to act against the gut microbes (see Supplemental Table S21). Thus, it is twice as much less likely that FDA-approved antibiotics would be toxic against gut microbes than antifungals.

3. Discussion

The identification of antimicrobial compounds assisted by machine-learning techniques has multiple advantages, such as reduction of the invested time to develop novel pharmaceuticals or to flag molecules that could have secondary antimicrobial activity [17]. An important aspect of these techniques is how to improve the reliability of these predictions. One way to achieve this is to increase the number of examples in the training and testing sets. In this work we propose that it is possible to use chemical compounds of different nature (peptides and NPCC) that are commonly modeled separately as antimicrobials to improve the reliability of the predictions. Here we show that indeed, the training sets that rendered the best classifiers of antimicrobial compounds were heterologous, those including NPCC and peptides (see Figures 1 and 2). We can compare our best classifier with previous works in terms of the learnability of our classes, that is, how well gut antimicrobial compounds are differentiated from non-antimicrobial gut compounds. In that sense, the numeric performance achieved by the best classifier on the testing set (AUC = 0.83) is comparable with the performance achieved with one of the best antimicrobial peptide classifiers (AUC = 0.85) recently reported [23], indicating that the learnability of heterologous training sets is as good as those of only peptides.

Another important aspect of our work is the molecular descriptors obtained to best classify gut antimicrobial compounds that included both peptides and NPCC. Although our goal was not to identify common descriptors for NPCC and peptides (these are already calculated by available packages, see Methods), we did look for those descriptors that are relevant to learn the difference between antimicrobials from non-antimicrobials. Our results indicate that the solution to this problem requires the transformation of 86 computed molecular descriptors, suggesting that other molecular descriptors, most likely associated to these 86 descriptors, may improve the current best-model performance.

In terms of improving the performance reported in this work, it is worth mentioning that we used peptides that were not tested by Maier et al. [7] yet, these peptides had reported antibiotic activity against at least one microorganism (*Escherichia coli*) found in the gut and tested by Maier and collaborators. On the other hand, the NPCC included in our work had antibiotic activity against at least one of the 40 gut microorganisms tested by Maier and collaborators. Hence, one alternative approach to improve the performance of classifiers aimed at identifying gut microorganisms would be to include antibiotics that target more common gut microorganisms; that would require further experimental data that is not currently available at present.

To the best of our knowledge, no previous machine-learning efforts to assist in the identification of broad-spectrum antibiotics have been reported; here the definition of broad-spectrum antibiotics was restricted to those acting against both pathogenic and non-pathogenic microorganisms. Hence, using a classifier trained to identify gut non-pathogenic antimicrobial compounds to predict this activity in FDA-approved antibiotics targeted against pathogenic microorganisms represents a way to identify broad-spectrum antibiotics. Our results suggest that half of the FDA-approved antibiotics are likely to have antimicrobial activity against the gut microorganisms indicating that these require further testing or investigation. For instance, two annotated narrow-spectrum antibiotics, amoxicillin and cephalexin, that were predicted to alter gut microbes are known to affect the gastrointestinal flora [24]. On the other hand, the broad-spectrum antibiotic ceftaroline fosamil recently approved by the FDA to treat bacterial pneumonia and skin infections, which was not predicted to affect the gut flora, was reported to have minor gastrointestinal effects during clinical trials [25].

How significant is our finding that almost half of the FDA-approved antibiotics are predicted to have a broad-spectrum activity? To address this question, we included in the discovery set a group of antifungal compounds. All microorganisms used to train our models were bacteria, hence we expected that these antifungals that act through a mechanism different from those reported for bacteria would be unlikely predicted to act against bacteria; lets refer to this negative prediction as *expectation-antifungal*. On the other hand, most FDA-approved antibiotics should unlikely present antibiotic activity against gut microbes, otherwise these would frequently have secondary gastrointestinal effects on patients; lets refer to this negative prediction as *expectation-FDA*. Then, to address the significance of our findings about broad-spectrum antibiotics requires evaluating *expectation-antifungal* and *expectation-FDA*; if FDA-approved drugs are less likely to act on gut microbes than antifungals then *expectation-FDA* < *expectation-antifungal*. Indeed, we observed that FDA-approved antibiotics are twice as much less likely to act against gut microbes than antifungals. Thus, our results indicate that even when FDA-approved antibiotics are safer (do not act against non-pathogenic resident gut bacteria) than our control group (antifungals), we identified some of these compounds that need to be re-assessed as potential promoters of resistance among microbes for their potential broad-spectrum activity.

In summary, we report a computational approach to use heterologous antimicrobial compounds (peptides and non-peptides) to improve the discriminatory power of machine-learning approaches. We show that training a classifier to identify antibiotics against the gut flora using heterologous training sets correctly anticipate adverse gastrointestinal reactions in patients receiving these antibiotics.

4. Materials and Methods

4.1. Materials

Peptides included in the training sets were obtained from the non-redundant data set of 20 public databases (see Table 5). Testing sets were derived from the work reported by Maier and collaborators (see Supplemental Tables S10–S18). Finally, a discovery set containing 750 FDA-approved drugs for treating human infectious diseases and 76 antifungal drugs was built from the ZINC database [26]. Molecular descriptors were computed with PadelDescriptor [27]. For every training and test set, we performed five different approaches to process the molecular descriptors for each peptide and/or NPCC. These included: no processing; eliminate every null value; substitute every "Infinity" value for 0 or 99,999,999; reduction of the dimensionality applying a principal component analysis implemented in WEKA package (see below). Since the substitution of Infinity values for 0 or 99,999,999 is not a conventional strategy, we performed an imputation of the Infinity and null values using the K nearest neighbor or mean imputation approaches, but only on the best model data set for comparison. That is, from the 9 training sets we generated a total of 45 training sets following the different approaches described before; the same applies to the 9 testing sets. For the discovery set only the transformation applied to the best classifier was performed.

Table 5. Antimicrobial peptide databases used in the present study.

Database	Focused on	Reference
BACTIBASE	Bacteriocins	[28]
Bagel	Bacteriocins	[29]
CAMP	General and Patented AMPs	[14]
DADP	Anuran AMPs	[30]
DAMPD	General AMPs *	[31]
DBAASP	General AMPs	[13]
Defensins	Defensins	[32]
HIPdb	Anti-HIV peptides	[33]
LAMP	General and Patented AMPs	[34]
MilkAMP	AMPs of dairy origin	[35]
PhytAMP	Plant AMPs	[36]
PenBase	Penaeidin AMPs	[37]
Peptaibol	Peptaibols	[38]
RAPD	Recombinant AMPs	[39]
AMPer	Eukaryotic AMPs	[40]
UniprotKb	General AMPs	[41]
YADAMP	General AMPs	[42]
AMSDb	Eukaryotic AMPs	[43]
APD	General AMPs	[44]
AVPdb	Antiviral peptides	[45]

* AMPs stands for Antimicrobial Peptides.

4.2. Method

To identify the best model to classify gut antimicrobial compounds, we followed a systematic method previously reported by our group [46]. Briefly, given the training sets, 52 different machine-learning algorithms implemented in WEKA [47] and their parameters were systematically analyzed to identify the algorithm, parameters and molecular descriptors that renders the lowest possible error in classification; this systematic analysis was performed by the Bayesian optimization algorithm implemented in AutoWEKA [48]. We ran AutoWEKA against any training set for 10, 90, 720, 2880 and 4320 minutes to identify when the optimization has reached a plateau in the classification error. Afterwards, a 10-fold cross validation and 67% split tests were performed in WEKA. Finally, these classifiers were evaluated against their corresponding testing sets. Two statistical parameters were chosen to evaluate the performance of the classifiers during the testing, including: Area under the ROC curve and correctly classified instances on the testing set. Therefore, a total of 5 statistical parameters were used to define the best classifiers, three for the training phase (adjusted estimated error rate on the training set; correctly classified instances in the training set after splitting 33% for testing; 10-fold cross-validation) and two for the testing phase (AUROC and correctly classified instances).

To identify the intersection set between the top 5 classifiers, we compared the predictions of these classifiers rendering 10 possible pairs of predictions on the discovery set; we used this set because not every classifier had the same testing set. The best model was identified using a combined score (see formula 1): the model with the lowest combined score was chosen. The model then was used to predict gut antimicrobial compounds in the discovery set using WEKA command line (see Supplemental File S1). To annotate as broad-spectrum or narrow-spectrum antibiotics, we used five different previous works that classified antibiotic action [22,49–52].

Supplementary Materials: The following are available online at http://bis.ifc.unam.mx:8080/ironbios/heteroml/, File S1: Script to execute the best model to predict antimicrobials on FDA-approved drugs, Table S1A–E: Training sets in ARF format for TrOnlyPeptides, Table S2A–E: Training sets in ARFF format for TrNPCC1, Table S3A–E: Training sets in ARFF format for TrNPCC2, Table S4A–E: Training sets in ARFF format for TrNPCC3, Table S5A–E: Training sets in ARFF format for TrNPCC4, Table S6A–L: Training sets in ARFF format for TrHeterologous1, Table S7A–E: Training sets in ARFF format for TrHeterologous2, Table S8A–E: Training sets in ARFF format for TrHeterologous3, Table S9A–E: Training sets in ARFF format for TrHeterologous4, Table S10A–E: Testing sets in ARF format for TeOnlyPeptides, Table S11A–E: Testing sets in ARF format for TeNPCC1, Table S12A–E: Testing sets in ARF format for TeNPCC2, Table S13A–E: Testing sets in ARF format for TeNPCC3, Table S14A–E: Testing sets in ARF format for TeNPCC4, Table S15A–J: Testing sets in ARF format for TeHetrelogous1, Table S16A–E: Testing sets in ARF format for TeHetrelogous2, Table S17A–E: Testing sets in ARF format for TeHetrelogous3, Table S18A–E: Testing sets in ARF format for TeHetrelogous4, Table S19: Parameter values for all models tested, Table S20: CScore values for all model tested, Table S21: Best models algorithms and corresponding parameters and Table S22: Discovery set.

Author Contributions: Conceptualization, G.D.R.; methodology, G.D.R., C.A.B. and R.A.N.L.; software, G.D.R. and R.A.N.L.; validation, R.A.N.L. and G.D.R.; formal analysis, R.A.N.L.; investigation, G.D.R. and R.A.N.L.; resources, G.D.R., C.A.B. and A.P.; data curation, R.A.N.L. and L.A.-M.; writing—original draft preparation, G.D.R.; writing—review and editing, C.A.B., A.P., L.A.-M. and R.A.N.L.; visualization, G.D.R. and R.A.N.L.; supervision, G.D.R. and C.A.B.; project administration, G.D.R.; funding acquisition, G.D.R.

Funding: This research was funded by the following agencies: CONACyT (gran numbers FOIN-219 and CB-252316), Programa de Apoyo a Proyectos de Investigación y Tecnológicos of the UNAM (grant numbers IG100416 and IN208014) and the Instituto de fisiología celular at UNAM. RN is enrolled as a graduate student of the Programa de maestría en ciencias bioquímicas at UNAM and received support from CONACyT with the fellowship number 631360.

Acknowledgments: To María Teresa Lara Ortiz for her technical assistance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Akova, M. Epidemiology of antimicrobial resistance in bloodstream infections. *Virulence* **2016**, *7*, 252–266. [CrossRef]
2. Aslam, B.; Wang, W.; Arshad, M.I.; Khurshid, M.; Muzammil, S.; Nisar, M.A.; Alvi, R.F.; Aslam, M.A.; Qamar, M.U.; Salamat, M.K.F.; et al. Antibiotic resistance: A rundown of a global crisis. *Infect. Drug Resist.* **2018**, *11*, 1645–1658. [CrossRef]

3. Roger, P.-M.; Montera, E.; Lesselingue, D.; Troadec, N.; Charlot, P.; Simand, A.; Rancezot, A.; Pantaloni, O.; Guichard, T.; Dautezac, V.; et al. Risk factors for unnecessary antibiotic therapy: A major role for clinical management. *Clin Infect Dis.* **2018**. [CrossRef] [PubMed]
4. Jackson, M.A.; Goodrich, J.K.; Maxan, M.-E.; Freedberg, D.E.; Abrams, J.A.; Poole, A.C.; Sutter, J.L.; Welter, D.; Ley, R.E.; Bell, J.T.; et al. Proton pump inhibitors alter the composition of the gut microbiota. *Gut* **2016**, *65*, 749–756. [CrossRef] [PubMed]
5. Rogers, M.A.M.; Aronoff, D.M. The influence of non-steroidal anti-inflammatory drugs on the gut microbiome. *Clin. Microbiol. Infect.* **2016**, *22*, 178.e1–178.e9. [CrossRef]
6. Flowers, S.A.; Evans, S.J.; Ward, K.M.; McInnis, M.G.; Ellingrod, V.L. Interaction between Atypical Antipsychotics and the Gut Microbiome in a Bipolar Disease Cohort. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* **2017**, *37*, 261–267. [CrossRef] [PubMed]
7. Maier, L.; Pruteanu, M.; Kuhn, M.; Zeller, G.; Telzerow, A.; Anderson, E.E.; Brochado, A.R.; Fernandez, K.C.; Dose, H.; Mori, H.; et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **2018**, *555*, 623–628. [CrossRef] [PubMed]
8. González-Díaz, H.; Prado-Prado, F.J.; Santana, L.; Uriarte, E. Unify QSAR approach to antimicrobials. Part 1: Predicting antifungal activity against different species. *Bioorg. Med. Chem.* **2006**, *14*, 5973–5980. [CrossRef]
9. Rath, E.C.; Gill, H.; Bai, Y. Identification of potential antimicrobials against Salmonella typhimurium and Listeria monocytogenes using Quantitative Structure-Activity Relation modeling. *PLoS ONE* **2017**, *12*, e0189580. [CrossRef]
10. Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F.J.; Salabert-Salvador, M.T.; Díaz-Villanueva, W.; Castro-Bleda, M.J.; Villanueva-Pareja, A. Artificial Neural Networks and Linear Discriminant Analysis: A Valuable Combination in the Selection of New Antibacterial Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031–1041. [CrossRef]
11. Nguyen, M.; Long, S.W.; McDermott, P.F.; Olsen, R.J.; Olson, R.; Stevens, R.L.; Tyson, G.H.; Zhao, S.; Davis, J.J. Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* **2018**. Available online: http://www.ncbi.nlm.nih.gov/pubmed/30333126 (accessed on 19 January 2019). [CrossRef] [PubMed]
12. Wang, Z.; Wang, G. APD: The Antimicrobial Peptide Database. *Nucleic Acids Res.* **2004**, *32*, D590–D592. [CrossRef] [PubMed]
13. Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H.L.; Squires, R.B.; Hurt, D.E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B.; et al. DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **2016**, *44*, D1104–D1112. [CrossRef] [PubMed]
14. Waghu, F.H.; Barai, R.S.; Gurung, P.; Idicula-Thomas, S. CAMP R3: A database on sequences, structures and signatures of antimicrobial peptides: Table 1. *Nucleic Acids Res.* **2016**, *44*, D1094–D1097. [CrossRef] [PubMed]
15. Del Rio, G.; Castro-Obregon, S.; Rao, R.; Ellerby, H.M.; Bredesen, D.E. APAP, a sequence-pattern recognition approach identifies substance P as a potential apoptotic peptide. *FEBS Lett.* **2001**, *494*, 213–219. [CrossRef]
16. Toropova, M.A.; Veselinović, A.M.; Veselinović, J.B.; Stojanović, D.B.; Toropov, A.A. QSAR modeling of the antimicrobial activity of peptides as a mathematical function of a sequence of amino acids. *Comput. Biol. Chem.* **2015**, *59*, 126–130. [CrossRef]
17. Durrant, J.D.; Amaro, R.E. Machine-Learning Techniques Applied to Antibacterial Drug Discovery. *Chem. Biol. Drug Des.* **2015**, *85*, 14–21. [CrossRef]
18. Battisti, A.; Zamuner, S.; Sarti, E.; Laio, A. Toward a unified scoring function for native state discrimination and drug-binding pocket recognition. *Phys. Chem. Chem. Phys.* **2018**, *20*, 17148–17155. [CrossRef]
19. Del Rio, G.; Koschützki, D.; Coello, G. How to identify essential genes from molecular networks? *BMC Syst. Biol.* **2009**, *3*, 102. [CrossRef] [PubMed]
20. Calahorra, M.; Sánchez, N.S.; Peña, A. Influence of phenothiazines, phenazines and phenoxazine on cation transport in Candida albicans. *J. Appl. Microbiol.* **2018**, *125*, 1728–1738. [CrossRef] [PubMed]
21. Acar, J. Broad- and narrow-spectrum antibiotics: An unhelpful categorization. *Clin. Microbiol. Infect.* **1997**, *3*, 395–396. [CrossRef]
22. Sarpong, E.M.; Miller, G.E. Narrow- and Broad-Spectrum Antibiotic Use among U.S. Children. *Health Serv. Res.* **2015**, *50*, 830–846. [CrossRef] [PubMed]

23. Beltran, J.A.; Aguilera-Mendoza, L.; Brizuela, C.A. Optimal selection of molecular descriptors for antimicrobial peptides classification: An evolutionary feature weighting approach. *BMC Genomics* **2018**, *19*, 672. [CrossRef]
24. NIH DailyMed. 26/November 2018. Available online: https://dailymed.nlm.nih.gov/dailymed/index.cfm (accessed on 19 January 2019).
25. File, T.M.; Wilcox, M.H.; Stein, G.E. Summary of Ceftaroline Fosamil Clinical Trial Studies and Clinical Safety. *Clin. Infect. Dis.* **2012**, *55*, S173–S180. [CrossRef] [PubMed]
26. Sterling, T.; Irwin, J.J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef]
27. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
28. Hammami, R.; Zouhir, A.; Le Lay, C.; Ben Hamida, J.; Fliss, I. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiol.* **2010**, *10*, 22. [CrossRef]
29. De Jong, A.; van Heel, A.J.; Kok, J.; Kuipers, O.P. BAGEL2: Mining for bacteriocins in genomic data. *Nucleic Acids Res.* **2010**, *38*, W647–W651. [CrossRef]
30. Novković, M.; Simunić, J.; Bojović, V.; Tossi, A.; Juretić, D. DADP: The database of anuran defense peptides. *Bioinformatics* **2012**, *28*, 1406–1407. [CrossRef] [PubMed]
31. Seshadri Sundararajan, V.; Gabere, M.N.; Pretorius, A.; Adam, S.; Christoffels, A.; Lehväslaiho, M.; Archer, J.A.C.; Bajic, V.B. DAMPD: A manually curated antimicrobial peptide database. *Nucleic Acids Res.* **2012**, *40*, D1108–D1112. [CrossRef]
32. Seebah, S.; Suresh, A.; Zhuo, S.; Choong, Y.H.; Chua, H.; Chuon, D.; Beuerman, R.; Verma, C. Defensins knowledgebase: A manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Res.* **2007**, *35*, D265–D268. [CrossRef]
33. Qureshi, A.; Thakur, N.; Kumar, M. HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides. *PLoS ONE* **2013**, *8*, e54908. [CrossRef] [PubMed]
34. Zhao, X.; Wu, H.; Lu, H.; Li, G.; Huang, Q. LAMP: A Database Linking Antimicrobial Peptides. *PLoS ONE* **2013**, *8*, e66557. [CrossRef] [PubMed]
35. Théolier, J.; Fliss, I.; Jean, J.; Hammami, R. MilkAMP: A comprehensive database of antimicrobial peptides of dairy origin. *Dairy Sci. Technol.* **2014**, *94*, 181–193. [CrossRef]
36. Hammami, R.; Ben Hamida, J.; Vergoten, G.; Fliss, I. PhytAMP: A database dedicated to antimicrobial plant peptides. *Nucleic Acids Res.* **2009**, *37*, D963–D968. [CrossRef] [PubMed]
37. Gueguen, Y.; Garnier, J.; Robert, L.; Lefranc, M.; Mougenot, I.; Lorgeril, J.; Janech, M.; Gross, P.S.; Warr, G.W.; Cuthbertson, B.; et al. PenBase, the shrimp antimicrobial peptide penaeidin database: Sequence-based classification and recommended nomenclature. *Dev. Comp. Immunol.* **2006**, *30*, 283–288. [CrossRef]
38. Whitmore, L.; Wallace, B.A. The Peptaibol Database: A database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.* **2004**, *32*, D593–D594. [CrossRef]
39. Li, Y.; Chen, Z. RAPD: A database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol. Lett.* **2008**, *289*, 126–129. [CrossRef]
40. Fjell, C.D.; Hancock, R.E.W.; Cherkasov, A. AMPer: A database and an automated discovery tool for antimicrobial peptides. *Bioinformatics* **2007**, *23*, 1148–1155. Available online: http://www.ncbi.nlm.nih.gov/pubmed/17341497 (accessed on 23 January 2019). [CrossRef]
41. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.
42. Piotto, S.P.; Sessa, L.; Concilio, S.; Iannelli, P. YADAMP: Yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents* **2012**, *39*, 346–351. [CrossRef] [PubMed]
43. Tossi, A.; Sandri, L. Molecular diversity in gene-encoded, cationic antimicrobial polypeptides. *Curr. Pharm. Des.* **2002**, *8*, 743–761. [CrossRef] [PubMed]
44. Wang, G.; Li, X.; Wang, Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087–D1093. [CrossRef]
45. Qureshi, A.; Thakur, N.; Tandon, H.; Kumar, M. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res.* **2014**, *42*, D1147–D1153. [CrossRef] [PubMed]

46. Corral-Corral, R.; Beltrán, J.; Brizuela, C.; Del Rio, G. Systematic Identification of Machine-Learning Models Aimed to Classify Critical Residues for Protein Function from Protein Structure. *Molecules* **2017**, *22*, 1673. [CrossRef]
47. Witten, I.H.; Ian, H.; Frank, E.; Hall, M.A.; Mark, A. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2011; 629p.
48. Kotthoff, L.; Thornton, C.; Hoos, H.H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
49. Newland, J.G.; Stach, L.M.; De Lurgio, S.A.; Hedican, E.; Yu, D.; Prasad, P.A.; Jackson, M.A.; Myers, A.L.; Zaoutis, T.E. Impact of a Prospective-Audit-With-Feedback Antimicrobial Stewardship Program at a Children's Hospital. *J. Pediatric Infect. Dis. Soc.* **2012**, *1*, 179–186. [CrossRef]
50. Newman, R.E.; Hedican, E.B.; Herigon, J.C.; Williams, D.D.; Williams, A.R.; Jason, G. Newland. Impact of a Guideline on Management of Children Hospitalized With Community-Acquired Pneumonia. *Pediatrics* **2012**, *129*, e597–e604. [CrossRef]
51. Di Pentima, M.C.; Chan, S. Impact of Antimicrobial Stewardship Program on Vancomycin Use in a Pediatric Teaching Hospital. *Pediatr. Infect. Dis. J.* **2010**, *29*, 707–711. [CrossRef]
52. Kreitmeyr, K.; von Both, U.; Pecar, A.; Borde, J.P.; Mikolajczyk, R.; Huebner, J. Pediatric antibiotic stewardship: Successful interventions to reduce broad-spectrum antibiotic use on general pediatric wards. *Infection* **2017**, *45*, 493–504. [CrossRef] [PubMed]

Sample Availability: All data used in this study are available as supplemental data.