Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning†

Christopher D. Fjell,*[,§,#] Håvard Jenssen,[#] Kai Hilpert,[#,▽] Warren A. Cheung,[§] Nelly Panté,[⊥] Robert E. W. Hancock,[#] and Artem Cherkasov[§]

Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia, 2733 Heather Street, Vancouver, British Columbia V5Z 3J5, Canada, Centre for Microbial Diseases and Immunity Research, University of British Columbia, 2259 Lower Mall, Vancouver, British Columbia V6T 1Z4, Canada, and Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada

Received December 5, 2008

The rise of antibiotic resistant pathogens is one of the most pressing global health issues. Discovery of new classes of antibiotics has not kept pace; new agents often suffer from cross-resistance to existing agents of similar structure. Short, cationic peptides with antimicrobial activity are essential to the host defenses of many organisms and represent a promising new class of antimicrobials. This paper reports the successful in silico screening for potent antibiotic peptides using a combination of QSAR and machine learning techniques. On the basis of initial high-throughput measurements of activity of over 1400 random peptides, artificial neural network models were built using QSAR descriptors and subsequently used to screen an in silico library of approximately 100,000 peptides. In vitro validation of the modeling showed 94% accuracy in identifying highly active peptides. The best peptides identified through screening were found to have activities comparable or superior to those of four conventional antibiotics and superior to the peptide most advanced in clinical development against a broad array of multiresistant human pathogens.

Introduction

Short cationic, amphipathic peptides that possess antimicrobial activity are present throughout the kingdoms of life. In the face of increasing antibiotic resistance in pathogenic microorganisms, such peptides have drawn significant attention as possible sources of novel antibacterial agents.[1−5] Although antimicrobial peptides (AMPs)[a] generally exhibit lower potency against susceptible bacterial targets compared to conventional low molecular weight antibiotic compounds, they hold several compensatory advantages including fast killing, broad range of activity, low toxicity, and minimal development of resistance in target organisms.[3,6]

The use of quantitative structure−activity relationships (QSAR) to predict antibacterial activity of peptides is a relatively recent development. QSAR analysis seeks to relate quantitative properties of a compound (known as descriptors) with other properties, such as drug-like activity or toxicity, and relies on physical properties that can be conveniently measured or calculated to predict in a nontrivial way other properties of interest such as biological activity. QSAR has become an integral part of screening programs in pharmaceutical drug discovery pipelines of small compounds and more recently in toxicological studies.[7] There are two aspects to QSAR analysis: the choice of the set of descriptors and the choice of statistical learning technique.

Previous QSAR analysis of antimicrobial peptides has been limited to comparisons between peptides with high similarity, for example, derivatives of lactoferricin[8−11] and protegrin and similar de novo peptides.[12−14] These QSAR studies have mainly utilized descriptors that are designed to model differences in properties of similar peptides, such as in the lactoferricin studies, or have used relatively simple descriptors such as charge, amphipathicity, and lipophilicity, the relationship of which has been demonstrated empirically from amino acid substitution studies.[13] Where larger sets of QSAR descriptors have been used, for example, for protegrin and analogues,[12,14] the models have been limited to linear models, resulting in only moderate predictive ability.

We decided to perform QSAR analysis on AMPs using a more intensive QSAR methodology that utilizes atomic scale molecular information, recently developed and applied to small molecules. We have recently reported that similar descriptors in combination with linear methods such as principal component analysis successfully predicted the activity of a library of highly related synthetic peptides, but failed to extrapolate to another library of dissimilar peptides.[15,16] These "inductive" QSAR descriptors (reviewed in ref 17) have been successfully applied to a number of molecular modeling studies including identification of the antibacterial activities of small compounds[18] and classification of antimicrobial compounds, conventional drugs, and drug-like substances, from an extensive set of over 2500 chemical structures, with up to 97% accuracy.[19] These studies have relied on modeling techniques of greater complexity than those previously applied to antimicrobial peptides. In particular, classification methods for such compounds were compared, including artificial neural networks (ANNs), *k*-nearest neighbors,

* Address correspondence to this author at the Centre for Microbial Diseases and Immunity Research, University of British Columbia, 2259 Lower Mall Research Station, Vancouver, BC V6T 1Z3, Canada [telephone (778) 384-5579; fax (604) 827-5566; e-mail cfjell@interchange.ubc.ca].

† The peptides described here have been submitted as part of a US patent application.

§ Division of Infectious Diseases, Department of Medicine.

Centre for Microbial Diseases and Immunity Research.

▽ Current Address: KIT (Karlsruhe Institute of Technology), Institute of Biological Interfaces, P.O.B. 3640, 76021 Karlsruhe, Germany.

⊥ Department of Zoology.

[a] Abbreviations: AMP, antimicrobial peptide; ANN, artificial neural network; AROC, area under the ROC curve; HPLC, high-pressure liquid chromatography; MIC, minimum inhibitory concentration; MRSA, methicillin-resistant *Staphylococcus aureus*; MS, mass spectroscopy; PCR, principal component regression; PLSR, partial least-squares regression; PPV, positive predictive value; QSAR, quantitative structure−activity relationship; Rel.IC_{50}, relative inhibitory concentration 50% (relative to control peptide Bac2A); ROC, receiver operating characteristics; SEM, scanning electron microscopy; TEM, transmission electron micrograph; VRE, vancomycin-resistant *Enterococcus*.

10.1021/jm8015365 CCC: $40.75
Published on Web 03/18/2009

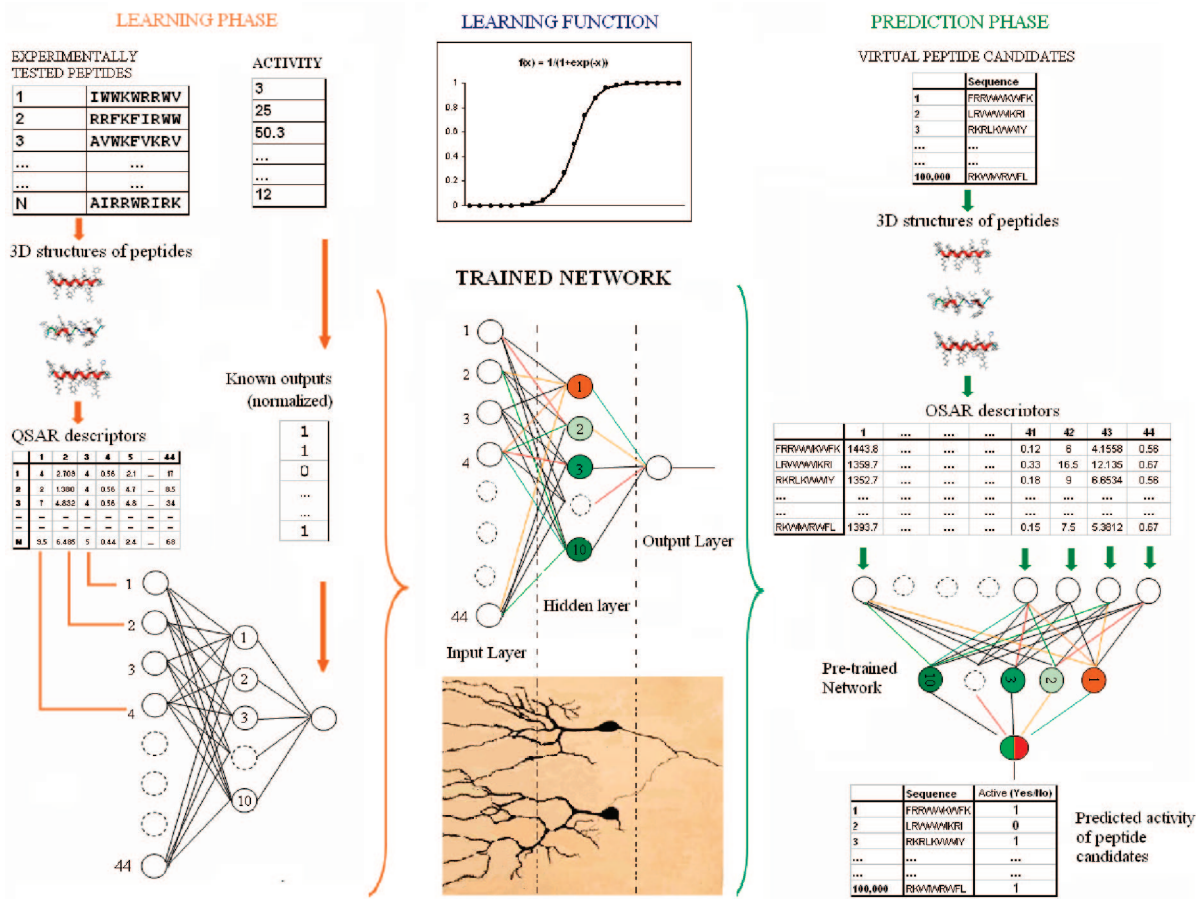

Figure 1. General workflow for QSAR modeling of antimicrobial peptides. The modeling proceeded in two phases, learning and prediction. In the learning phase (left column), the 3D structures of experimentally tested peptides were estimated, and the QSAR descriptors were calculated on the basis of these structures. Artificial neural networks (ANNs) were trained to classify peptides as active or inactive on the basis of the descriptor values and experimental activity level. ANNs involve a nonlinear transformation of the inputs and are organized in three layers, with connections and function inspired by natural neuron connections (center column). In the prediction phase (right column), the activities of virtual peptides were predicted using calculated descriptor values (based on estimated 3D structures) as input to the trained ANNs.

linear discriminative analysis, and multiple linear regression, and it was found that ANNs result in generally more accurate predictions for classification, followed closely by *k*-nearest neighbors methods.[20]

These higher complexity models use a larger number of parameters and therefore require greater amounts of data. These data were available from the recently developed high-throughput method for screening large numbers of peptides for antibacterial activity.[21] This method rapidly synthesizes peptides on cellulose support to create peptides that are not limited in sequence diversity. The peptides have been assayed for antimicrobial activity using a strain of *Pseudomonas aeruginosa* engineered to constitutively luminesce due to an incorporated five-gene luciferase cassette. By measuring the decrease in luminescence due to killing and loss of ability to energize luminescence in the bacteria, a large number of peptides were screened for antibacterial activity in an automated manner.

In the current work, we applied the methods of atomic resolution QSAR combined with complex, nonlinear modeling to accurately predict the antibacterial activities of short cationic peptides containing high sequence diversity. By combining high-throughput generation of synthetic peptides with a high-throughput antibacterial assay, we were able to apply these methods to a larger data set of peptides than has been used to date. We demonstrate that this combination of experimental procedure and QSAR analysis provides dramatic improvement in prediction of diverse antibacterial peptides. With methods we describe here, we have performed an efficient, large-scale in silico screening for antibacterial peptides that has yielded several potential drug leads. We reported elsewhere a summary version of these methods[22] and, in particular, extensive in vitro and in vivo results with some of the better peptides; we report the details of the methodology here.

Results and Discussion

The overall process used for QSAR modeling of antimicrobial peptides is shown in Figure 1. In summary, the starting point was a set of random peptides with measured activity. For these peptides the 3D structure was estimated and used to calculate QSAR descriptors for each. Models for peptide activity were built using artificial neural networks based on these descriptors and the known levels of activity. These models were then used to computationally assess a much larger set of virtual peptides for predicted activity. The accuracy of the predictions was independently assessed by synthesizing and testing many peptides with various levels of predicted activity.

Effect of Control Antibacterial Peptide on Bacteria. The effect of treatment of *P. aeruginosa* with the active control peptide Bac2A is shown in transmission electron micrographs (TEMs) (Figure 2, left side). These electron micrographs showed that Bac2A had a dramatic effect on the morphology of the bacterial cell surface. Whereas the cell surface of control untreated bacteria appeared to be smooth (see Figure 2A, left

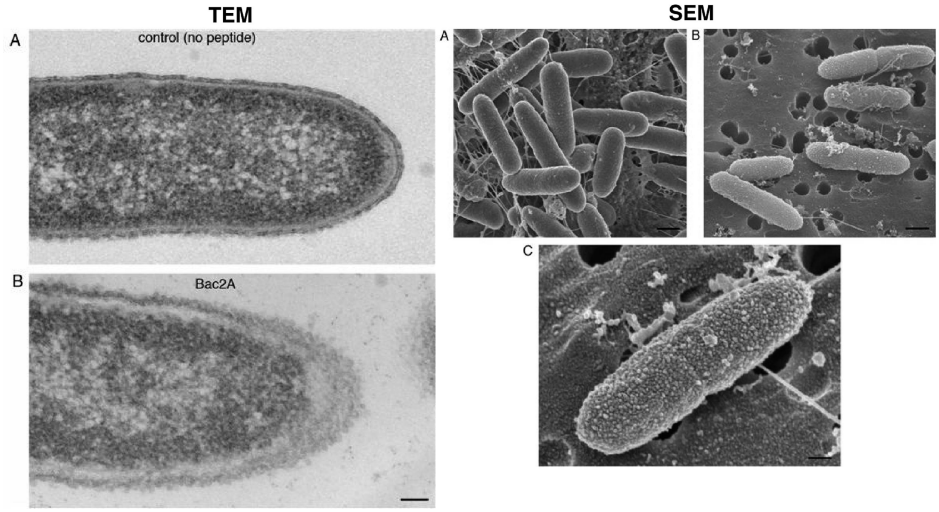

Figure 2. Transmission electron micrographs (TEM) and scanning electron micrographs (SEM) of *P. aeruginosa*. TEM and SEM are shown for control untreated (TEM A, SEM A) and Bac2A-treated (TEM B, SEM B,C). Bac2A concentration was at the MIC. Bacteria were incubated with Bac2A for 1 h at 37 °C before fixation and preparation for TEM or SEM. Scale bar is 100 nm (TEM A and B, SEM C) or 500 nm (SEM A and B).

side), the Bac2A-treated bacteria had cell surfaces that were severely damaged and contained numerous blebs (Figure 2B, left side), a well-known phenomenon observed when bacterial cells are exposed to cationic peptides.[23] In addition, the space between the cell wall and plasma membrane appeared to be swollen. The blebs of the cell wall were better appreciated when the surface of Bac2A-treated bacteria were visualized by scanning electron microscopy (SEM) (Figure 2, right side).

Peptide Data Sets for Model Training. Two initial sets of synthetic peptides of nine amino acids in length were assayed for antibacterial activity. Set A consisted of 933 peptides; set B consisted of 500 peptides. The primary sequences of set A were chosen with a bias toward enrichment of these sets for the amino acid proportions of our previously isolated peptides with antibacterial activity based on previous studies.[21,24] Subsequently, set B peptides were designed with the adjusted amino acid compositions of the initial peptide population plus set A peptides, as shown in Supporting Information Supplementary Figure 1. In both sets, there were no constraints on the amino acid proportions found within any particular peptide. The two sets were progressively prepared by synthesis on a cellulose support and assayed for activity against *P. aeruginosa* using a luciferase reporter assay as described previously.[21]

Calculation of Peptide Activity. Peptide antibacterial activity was measured using the luminescence assay, which assesses the loss of energy generation capacity shown with antimicrobial peptides to proportionately reflect lethality as previously described.[21,24] Peptides were assayed in a dilution series with relative IC_{50} (rel IC_{50}) values of the experimental peptides determined by curve fitting and parameter estimation (for illustration, see Figure 3) as described below. The fit of the luminescence experimental values was generally good except for peptides with very low activity, for which the plateau at low luminescence (higher killing concentrations) was not present. These inactive peptides were assigned the rel IC_{50} value of 25. The activity of the two sets is shown in Table 1 (training set A and B rows) classified into higher activity (rel $IC_{50} < 50\%$ of the control peptide, Bac2A), similar activity (rel IC_{50} between 50 and 150% of control), and lower activity (rel $IC_{50} > 150\%$ of control).

QSAR Descriptors and Model Building. A large number of QSAR descriptors are available to describe the physical

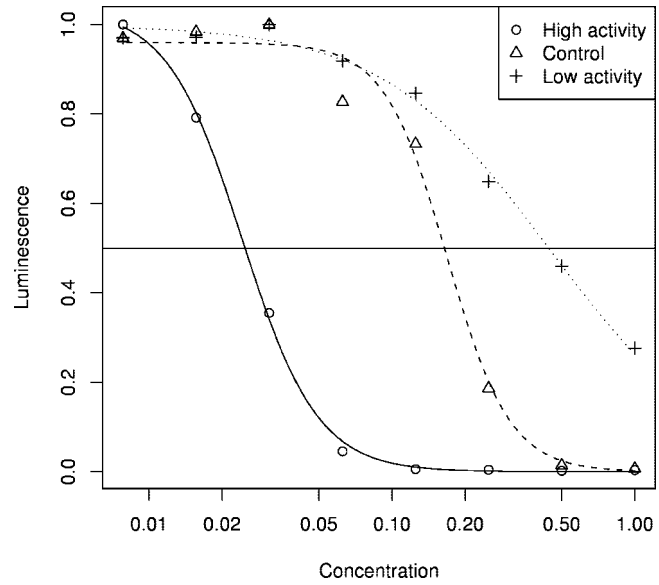

Figure 3. Luminescence profile of a dilution series for three peptides-.The luminescences for three peptides having high, medium (control peptide), and low activity are shown. Luminescence and concentrations were scaled to maximum of 1.0. The *X*-axis value equivalent to the point where the horizontal line at luminescence of 0.5 crossed the fitted curves indicated the relative IC_{50} value for each peptide.

chemistry of compounds. A total of 77 descriptors were calculated here for each peptide using an estimated structural conformation based on energy minimization in the gas phase (using MMFF94 force field). We have found that performance of predictions of activity is insensitive to the method used to estimate the three-dimensional structure; performance was similar for peptides with structures using straight backbone and structures estimated by energy minimization in the gas phase or an implicit solvent (Supporting Information Supplementary Table 2). Some descriptor values were found to be highly correlated with each other, which led to problems in modeling; therefore, a set of 44 descriptors was chosen that showed <95% correlation to any other selected descriptor. All descriptors are shown in Supporting Information Supplementary Table 1; those used for modeling are indicated. Descriptors with large cor-

Table 1. Activities of Peptides from Training Sets and Quartiles in the 100,000 Test Set[a]

	rel IC_{50}			
data set	higher activity (<0.5)	similar activity (0.5−1.5)	lower activity (>1.5)	median
set A	35 (3.8%)	210 (22.5%)	688 (73.7%)	2.12
set B	14 (2.8%)	114 (22.8%)	372 (74.4%)	3.33
Q1	47 (94%)	2 (4%)	1 (2%)	0.23
Q2	32 (64%)	15 (30%)	4 (8%)	0.35
Q3	1 (2%)	5 (10%)	44 (88%)	4.38
Q4	0 (0%)	0 (0%)	50 (100%)	8.34

[a] Numbers of peptides with various levels of antibacterial activity are shown. Q1, top of 1st quartile; Q2, top of 2nd quartile; Q3, bottom of 3rd quartile; Q4, bottom of 4th quartile. Rel IC_{50} is the relative IC_{50}, the ratio of the IC_{50} for the experimental peptide to the IC_{50} of Bac2A. Peptides of which the highest concentration failed to reduce the luminescence by at least 50% were identified as inactive.

relations (>0.7 or <−0.7) are shown in Supporting Information Supplementary Table 3.

We performed exploratory data analysis using principal component regression (PCR) and partial least-squares regression (PLSR) on set A data in an attempt to identify significant descriptors. PCR and PLSR were initially performed using 44 principal components (the maximum possible number, equal to the number of descriptors). The fraction of activity that can be accounted for by models with increasing numbers of components (cross-validated R^2) is shown in Supporting Information Supplementary Figures 2 and 3. R^2 is maximum at 0.12 for PLSR at six components, whereas R^2 is maximum (also at 0.12) using all 44 components for PCR. The best PLSR model, using six components, fit the data poorly (Supporting Information Supplementary Figure 4). The loadings on the first three PLSR components (the contributions of each descriptor to each component) are shown in Supporting Information Supplementary Figure 5. There are no descriptors that clearly dominate the model. The largest absolute loading on component 1 is 0.25 (+0.25 for vdw_area and −0.25 for Average_hardness), but 21 descriptors have absolute values above 0.15 (Supporting Information Supplementary Table 4). These results suggest that modeling using PCR or PLSR cannot capture important features of the data.

We used ANNs (Supporting Information Supplemental Figure 7) to model antibacterial activity because this has already been successfully applied to small molecules (for example, see ref 18). ANNs typically rank highly among machine learning techniques in predictive performance, and, in addition, they are relatively insensitive to the presence of noise and correlated inputs. We used a network configuration with one hidden layer of 10 nodes, 44 input nodes (one for each descriptor), and 1 output node. A variety of other network configurations were also evaluated with the number of hidden nodes per hidden layer (network length) varying between 1 and 20 and the number of hidden node layers (network width) varying from 1 to 3. Prediction performance was calculated as positive predictive value [PPV, (true positive)/(true positive + false positive)]. Except for network lengths of <8 with a width of 1, there did not appear to be any advantage of using larger numbers of hidden layers or more nodes per hidden layer than around 10 (Supporting Information Supplementary Figure 6).

Validation of Model Performance. We assessed the ability of the ANN models to predict antibacterial activity by first classifying the top 5% of the set A and B peptides as active according to the rel IC_{50} values—this corresponded to an approximate rel IC_{50} threshold of 0.6 (0.61 for set A and 0.56 for set B). A 10-fold cross-validation was performed as described below with 90% of data allocated to training and 10% to validation (i.e., reserving a different 10% for each of the 10 validation studies). Sets A and B were synthesized and assayed at different times as described above, and some systematic differences were observed in the luminescence results for peptides of very low and very high activity. Therefore, we treated sets A and B separately, in addition to combining them into pooled set, set A+B. The performance of the three models was assessed using receiver operating characteristics (ROC) curves (Supporting Information Supplementary Figure 8) and the area under the ROC curves (AROCs). AROC values approaching 1 indicate an increasing ability to accurately classify data; AROC values close to 0.5 indicate a poor ability to classify. The average AROC value for sets A and B and the combined set A+B were found to be (mean ± standard deviation, SD) 0.87 ± 0.10, 0.83 ± 0.12, and 0.80 ± 0.09, respectively. These data show that the cross-validated performance of the models to predict peptide activity was quite good. We integrated the large number of models generated during the cross-validation in a consensus approach to allow a combined, single prediction for a given peptide. We did this using a "voting" system, whereby each of the 20 models (10 each for set A, set B, and the combined set A+B) was used to rank a set of test peptides. If a peptide was ranked in the top 5% of the set, it received one vote by the model. Because we used 30 models, a peptide could receive up to 30 votes. When different peptides received the same number of votes, these peptides were further ranked using the average ranking produced by the 30 models. (For example, a peptide receiving 20 votes with an average ranking of 500 in a list of 100,000 peptides would be ranked higher than another peptide with 20 votes and an average ranking of 600.)

Independent Model Testing. To perform an independent assessment of this approach to identify highly active antibacterial peptides, we created a random set of approximately 100,000 peptides in an independent test set using the same global amino acid proportions as set B (see Supporting Information Supplementary Figure 1). When we calculated the 44 QSAR descriptors for each peptide, a modest number of peptides yielded values that were >15% outside the range of descriptor values encountered in sets A and B and were not considered further, because it is believed that this would lead to less reliable performance of the models. This left a total of 99,577 test peptides. Each of these peptides was ranked numerically using a voting system as described below. Because these models were intended to classify peptides as active or inactive, rather than to predict actual activity levels, the ranked list of test peptides indicated the likelihood that a peptide is highly active. Interestingly despite the very large difference in predicted activities, the peptides in each quartile had rather similar bulk physical properties (charge, hydrophobicity, hydrophobic moment) (Table 1; Supporting Information Supplementary Table 5), indicating the importance of using a broad variety of descriptors in neural network assisted modeling.

To independently evaluate these predictions of peptide activity, we selected and synthesized a total of 200 candidate peptides comprising sets of 50 candidate peptides at four positions of ranking. Quartile 1 (Q1) peptides were ranked in the topmost 50 positions and considered to be the most likely to be more active than control. Quartile 2 (Q2) peptides were ranked at the start of the second quartile, positions 24,895−24,944, and thus considered likely to be more active than control. Quartile 3 (Q3) peptides were ranked at the end of the third quartile, positions 74,633−74,682, and considered likely to be less active than control. Quartile 4 (Q4) peptides were ranked

Table 2. Predicted Activity Rank and Experimental IC_{50} Values for Selected Test Peptides[a]

peptide rank	quartile	sequence	cumulative vote	measured rel IC_{50}	charge	hydrophobic fraction	hydrophobic moment
1	1	RWRWKRWWW	29	0.25	4	0.56	1.48
2	1	RWRRWKWWW	29	0.40	4	0.56	1.96
3	1	RWWRWRKWW	29	0.28	4	0.56	2.11
8	1	KIWWWWRKR	27	0.13	4	0.56	2.06
9	1	RWRRWKWWL	27	0.08	4	0.56	2.12
10	1	KRWWKWIRW	27	0.04	4	0.56	4.65
20	1	WRWWKIWKR	26	0.14	4	0.56	4.8
36	1	KRWWKWWRR	25	0.13	5	0.44	5.9
45	1	WKRWWKKWR	25	0.20	5	0.44	4.7
48	1	WKKWWKRRW	25	0.19	5	0.44	2.4
24,895	2	IRMWVKRWR	0	0.61	4	0.56	4.24
24,896	2	RIWYWYKRW	0	0.36	3	0.67	4.06
24,897	2	FRRWWKWFK	0	0.12	4	0.56	5.40
24,901	2	LRWWWIKRI	0	0.33	3	0.67	0.99
24,910	2	RKRLKWWIY	0	0.18	4	0.56	2.0
24,913	2	KKRWVWIRY	0	0.22	4	0.56	1.0
24,915	2	KWKIFRRWW	0	0.16	4	0.56	3.5
24,919	2	RKWIWRWFL	0	0.15	3	0.67	2.8
24,921	2	IWWKWRRWV	0	0.29	3	0.67	3.5
24,944	2	RRFKFIRWW	0	0.24	4	0.44	2.1
74,655	3	AVWKFVKRV	0	8.18	3	0.67	4.5
74,658	3	AWRFKNIRK	0	9.20	4	0.44	1.8
74,665	3	KRIMKLKMR	0	6.50	5	0.44	4.0
74,673	3	KIRRKVRWG	0	10.55	5	0.33	2.02
74,674	3	AIRRWRIRK	0	4.62	5	0.44	5.94
74,675	3	WRFKVLRQR	0	7.08	4	0.44	4.20
74,677	3	FMWVYRYKK	0	1.51	3	0.67	1.81
74,678	3	RGKYIRWRK	0	3.83	5	0.33	4.94
74,679	3	WVKVWKYTW	0	5.64	2	0.67	2.41
74,680	3	VVLKIVRRF	0	25.00	3	0.67	1.86
99,568	4	GRIGGKNVR	0	9.12	3	0.22	4.30
99,569	4	NKTGYRWRN	0	8.33	3	0.22	2.75
99,570	4	VSGNWRGSR	0	8.54	2	0.22	2.67
99,571	4	GWGGKRRNF	0	7.38	3	0.22	1.13
99,572	4	KNNRRWQGR	0	6.45	4	0.11	2.88
99,573	4	GRTMGNGRW	0	6.93	2	0.22	1.40
99,574	4	GRQISWGRT	0	8.04	2	0.22	1.94
99,575	4	GGRGTRWHG	0	8.60	3	0.11	2.63
99,576	4	GVRSWSQRT	0	8.50	2	0.22	2.56
99,577	4	GSRRFGWNR	0	8.10	3	0.22	0.58

[a] Forty peptides are shown from the 200 total candidate peptides. Hydrophobic moment is given using the Eisenberg scale.

at the end of the fourth quartile, positions 99,528−99,577, and considered most likely to be less active than control. These 200 predicted peptides were synthesized and assayed for activity using the *lux* assay. As summarized in Table 1, the activity was predicted very accurately by the system. Of the 50 peptides in the most likely active set (Q1), 94% were found to be more active than control. Of the set considered less likely to be active (Q2), 64% were better than control. Of the peptides predicted to be much less active (Q3), 88% had lower activity than control. In the set considered least likely to be active (Q4), all (100%) were less active than control. All 200 candidate peptides are shown in the Supporting Information Supplementary Table 5 along with the rank, cumulative vote, experimentally determined rel IC_{50} values, and selected physical properties (charge, hydrophobic fraction, and hydrophobic moment).

Ten peptides from each quartile are shown in Table 2 to permit discussion. Consistent with the bulk features of the entire library of sequences, for these peptides the charge and hydrophobicity showed a large degree of overlap for most quartiles. Only certain of the peptides from Q4 showed a noticeable difference in these physical properties, specifically in showing a lower charge and hydrophobicity. The importance of charge, hydrophobicity, and amphipathicity for antibacterial activity of peptides is well-known.[6,25] However, in these groups of peptides there was a clear difference only between the most active and least active sets (Q1 and Q4) in terms of charge and hydrophobicity, whereas the differences in activity across all quartiles were quite dramatic. A graphic example that these properties are by themselves insufficient to make predictions can be observed by comparing peptides 10 and 74,675 that have very similar values for charge (+4), hydrophobicity (0.44−0.56), and hydrophobic moment (a measure of amphipathicity; 4.2−4.65), but have rel IC_{50} values that differ by >100-fold (0.04 and 7.1). This demonstrates that the success in predictions is not based on identifying potent peptides using previously known characteristics.

Antibacterial Activity of Predicted Peptides against Resistant Strains. A selection of 18 of these 200 peptides was synthesized in bulk and tested against a large variety of drug-resistant bacterial pathogens (Table 3). A total of 13 peptides from quartiles 1 and 2 with high activity and 5 peptides from quartile 3 with low assayed activity were evaluated for their in vitro effect (MIC activity) against several of the most multidrug-resistant and problematic pathogens including strains of multidrug-resistant *P. aeruginosa*, methicillin-resistant *Staphylococcus aureus* (MRSA), *Enterobacter cloacae* with derepressed chromosomal β-lactamase, extended spectrum β-lactamase producing *Escherichia coli* and *Klebsiella pneumonia*, and vancomycin-resistant *Enterococcus faecalis* and *Enterococcus faecium* (VRE). Peptides from the first and second quartiles had

Table 3. Activities against Multiresistant Superbugs of Selected Peptides Predicted through QSAR Analysis Compared to the Peptide Bac2A[a]

peptide		MIC (μM)																				
		P. aeruginosa							*P. maltophilia*	*E. cloacae*	*E. coli*		*K. pneumonia*		*S. aureus*		*E. faecalis*			*E. faecium*	
ID	sequence	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Bac2A	RLARIVVIRVAR	48 (34)	192 (135)	95 (67)	192 (135)	95 (67)	95 (67)	12 (8.4)	3.0 (2.1)	24 (17)	24 (17)	24 (17)	192 (135)	192 (135)	24 (17)	24 (17)	12 (8.4)	48 (34)	48 (34)	12 (8.4)	3.0 (2.1)
8	KIWWWWRKR	5.9 (4.1)	47 (32)	24 (17)	47 (32)	47 (32)	12 (8.3)	5.9 (4.1)	3.0 (2.1)	94 (65)	12 (8.3)	5.9 (4.1)	189 (130)	47 (32)	5.9 (4.1)	5.9 (4.1)	24 (17)	94 (65)	94 (65)	5.9 (4.1)	1.5 (1)
9	RWRRWKWWL	2.9 (2.0)	12 (8.1)	12 (8.1)	23 (16)	5.8 (3.9)	12 (8.1)	0.3 (0.2)	0.7 (0.5)	5.8 (3.9)	5.7 (3.9)	2.9 (2.0)	46 (31)	11 (7.4)	2.9 (2)	2.9 (2)	23 (16)	92 (62)	92 (62)	5.7 (3.9)	1.4 (0.9)
20	WRWWKIWKR	5.9 (4.1)	24 (17)	24 (17)	47 (32)	12 (8.3)	47 (32)	1.5 (1.0)	0.8 (0.6)	12 (8.3)	5.9 (4.1)	5.9 (4.1)	94 (65)	24 (17)	3 (2.1)	3 (2.1)	24 (17)	94 (65)	94 (65)	5.9 (4.1)	1.5 (1.0)
45	WKRWWKKWR	23 (16)	46 (31)	46 (31)	93 (63)	23 (16)	46 (31)	5.8 (4.0)	1.4 (1.0)	93 (63)	23 (16)	2.9 (2.0)	186 (127)	46 (31)	5.8 (4.0)	5.8 (4.0)	93 (63)	>186 (127)	>186 (127)	23 (16)	5.8 (4.0)
48	WKKWWKRRW	23 (16)	46 (31)	46 (31)	93 (63)	46 (31)	46 (31)	5.8 (4.0)	1.4 (1.0)	23 (16)	23 (16)	2.9 (2.0)	186 (127)	46 (31)	1.4 (1.0)	2.9 (2.0)	93 (63)	>186 (127)	>186 (127)	12 (8.2)	5.8 (4.0)
24,897	FRRWWKWFK	1.5 (1.0)	12 (8.3)	3.0 (2.1)	5.9 (4.1)	5.9 (4.1)	24 (17)	1.5 (1.0)	0.8 (0.6)	24 (17)	12 (8.3)	6.1 (4.2)	97 (67)	24 (17)	1.5 (1.0)	3 (2.1)	24 (17)	195 (135)	97 (67)	6.1 (4.2)	6.1 (4.2)
24,901	LRWWWIKRI	13 (9.6)	50 (37)	25 (18)	50 (37)	25 (18)	25 (18)	6.3 (4.6)	3.2 (2.4)	50 (37)	25 (18)	13 (9.6)	201 (148)	50 (37)	6.3 (4.6)	6.3 (4.6)	13 (9.6)	50 (37)	25 (18)	6.3 (4.6)	1.5 (1.1)
24,910	RKRLKWWIY	25 (18)	50 (37)	50 (37)	50 (37)	50 (37)	50 (37)	6.3 (4.7)	3.2 (2.4)	13 (9.6)	6.3 (4.7)	6.3 (4.7)	>202 (149)	50 (37)	6.3 (4.7)	6.3 (4.7)	50 (37)	>202 (149)	202 (149)	13 (9.6)	3.2 (2.4)
24,913	KKRWVWIRY	25 (19)	51 (38)	25 (19)	51 (38)	25 (19)	25 (19)	3.2 (2.4)	1.6 (1.2)	25 (19)	25 (19)	13 (9.7)	204 (152)	51 (38)	13 (9.7)	13 (9.7)	25 (19)	102 (76)	102 (76)	6.4 (4.8)	6.4 (4.8)
24,915	KWKIFRRWW	12 (8.5)	24 (17)	24 (17)	48 (34)	12 (8.5)	48 (34)	3.1 (2.2)	1.5 (1.1)	6.0 (4.3)	24 (17)	12 (8.5)	194 (138)	97 (69)	3.1 (2.2)	3.1 (2.2)	24 (17)	97 (69)	97 (69)	24 (17)	6.1 (4.3)
24,919	RKWIWRWFL	6.1 (4.4)	12 (8.6)	3.1 (2.2)	6.1 (4.4)	6.1 (4.4)	3.1 (2.2)	1.5 (1.1)	1.5 (1.1)	3.0 (2.2)	1.5 (1.1)	1.5 (1.1)	3.1 (2.2)	3.1 (2.2)	3.1 (2.2)	3.1 (2.2)	6.1 (4.4)	24 (17)	24 (17)	3.1 (2.2)	3.1 (2.2)
24,921	IWWKWRRWV	6.0 (4.2)	48 (34)	12 (8.5)	48 (34)	12 (8.5)	48 (34)	6.0 (4.2)	1.5 (1.1)	6.0 (4.2)	24 (17)	6.0 (4.2)	96 (68)	12 (8.5)	3 (2.1)	3 (2.1)	24 (17)	48 (34)	48 (34)	6 (4.2)	3 (2.1)
24,944	RRFKFIRWW	6.1 (4.4)	24 (17)	49 (35)	49 (35)	12 (8.6)	49 (35)	3.1 (2.2)	0.8 (0.6)	12 (8.6)	12 (8.6)	12 (8.6)	98 (70)	49 (35)	6.1 (4.4)	6.1 (4.4)	12 (8.6)	98 (70)	49 (35)	6.1 (4.4)	6.1 (4.4)
74,655	AVWKFVKRV	240 (211)	>240 (211)	240 (211)	>240 (211)	240 (211)	>240 (211)	120 (106)	60 (53)	>240 (211)	240 (211)	120 (106)	>240 (211)	>240 (211)	240 (211)	240 (211)	>240 (211)	>240 (211)	>240 (211)	120 (106)	120 (106)
74,658	AWRFKNIRK	>223 (182)	>223 (182)	>223 (182)	>223 (182)	>223 (182)	>223 (182)	111 (91)	>223 (182)	>223 (182)	>223 (182)	223 (182)	>223 (182)	>223 (182)	223 (182)	>223 (182)	>223 (182)	>223 (182)	>223 (182)	>223 (182)	223 (182)
74,665	KRIMKLKMR	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)	>226 (187)
74,674	AIRRWRIRK	>217 (172)	>217 (172)	>217 (172)	>217 (172)	>217 (172)	>217 (172)	217 (172)	108 (86)	108 (86)	108 (86)	108 (86)	>217 (172)	>217 (172)	54 (43)	54 (43)	>217 (172)	>217 (172)	>217 (172)	108 (86)	14 (11)
74,680	VVLKIVRRF	>241 (213)	>241 (213)	>241 (213)	>241 (213)	>241 (213)	>241 (213)	241 (213)	60 (53)	241 (213)	>241 (213)	241 (213)	>241 (213)	>241 (213)	241 (213)	241 (213)	241 (213)	>241 (213)	>241 (213)	241 (213)	60 (53)

[a] Peptides from the top quartile (8−48) were compared to peptides from the 2nd (24,897−24,944) and 3rd (74,655−74,680) quartiles. Columns headings: peptide ID indicates the control Bac2A or the test peptide by rank number. Columns give MIC values in μM and μg/mL in parentheses, measured in 3−5 replicates for A, *P. aeruginosa* wild type strain H103; B, C, D, *P. aeruginosa* multidrug-resistant strains from Brazil 9, 198, and 213, respectively; E, F, G, *P. aeruginosa* Liverpool epidemic strains LES400, H1030, and H1027, respectively; H, *P. maltophilia* ATCC13637; I, constitutive class C chromosomal β-lactamase expressing *E. cloacae* 218R; J, K, extended-spectrum β-lactamase-producing (ESBL) *E. coli* (clinical strains 63103 and 64771); L, M, ESBL-resistant *K. pneumonia* (clinical strains 61962 and 63575); N, *S. aureus* ATCC25923; O, methicillin-resistant *S. aureus* strain C623; P, *E. faecalis* ATCC29212; Q, R, vancomycin-resistant *E. faecalis* [clinical isolates w61950 (VanA) and f43559 (VanB)]; S,T, vancomycin-resistant *E. faecium* [clinical isolates mic80 (VanA) and t62764 (VanB)].

significant in vitro inhibitory activity against antibiotic-resistant bacteria. Moreover, some peptides from the first quartile, such as 8 and 9, exhibited MICs of 0.3−10 μM against most of the tested "superbugs", compared to the only antimicrobial peptide to show efficacy to date in advanced clinical trials, MX-226,[3] which exhibited MICs of 10−76 μM.[22] These results characterize the developed peptides as excellent antibiotic candidates for treating some of the most recalcitrant and dangerous human infections. As reported elsewhere,[22] two other peptides identified from the first quartile were also found to be protective against *S. aureus* infection in animal models.

Materials and Methods

Electron Microscopy of AMPs. TEM micrographs of thin sections of *P. aeruginosa* were untreated and treated with Bac2A (sequence RLARIVVIRVAR) at the MIC (50 μg/mL) for 1 h at 37 °C. For control, bacteria were mock incubated and prepared for embedding/thin section electron microscopy in the same way as the peptide-treated bacteria. SEM micrographs of *P. aeruginosa* were prepared for control untreated and Bac2A-treated (50 μg/mL). Bacteria were incubated with Bac2A for 1 h at 37 C before fixation and preparation for SEM.

Empirical Peptide Sequences. Two experimental sets of peptides were created, one consisting of 933 peptides (set A) and another with 500 peptides (set B). Peptide sequences in these sets were selected randomly on the basis of the amino acid composition (Supporting Information Supplementary Figure 2) of the most active peptides from prior experiments. The amino acid proportions for set A were determined on the basis of previous amino acid substitution studies,[21] and proportions for set B were adjusted on the basis of the early analysis of set A activities. For modeling (described below) three training sets were prepared, consisting of the set of 933 peptides (set A), the set of 500 peptides (set B), and a set created from combining the 933 and 500 peptide sets (set A+B).

A set of 100,000 random peptide sequences were generated with the same amino acid proportions as used for set B. A total of 311 peptides were removed because they were either duplicates or contained QSAR descriptors that differed by >15% from the range of those assessed for the training sets, leaving 99,577 peptides (the test set). Peptides from this set were evaluated in silico, and 200 (50 from each quartile) were selected for synthesis and assay.

Peptide SPOT Synthesis and Screening. SPOT synthesis of peptides in arrays was performed as previously described.[21,26] Briefly, peptides were synthesized on cellulose support with a pipetting robot using two glycine residues as linker. Peptides were cleaved from the dried membrane in an ammonia atmosphere, resulting in free peptides with two glycines at the amidated C terminus due to the linker sequence. The peptide spots were punched out and transferred to 96-well microtiter plates in sets of 10 along with a positive control peptide (Bac2A) and an unrelated peptide (GATPEDLNQKLS) or an empty well for negative control. An overnight culture of *P. aeruginosa* strain H1001 (*fliC::luxCDABE*) was diluted at 1:500 ratio with 100 mM Tris-HCl buffer (pH 7.3), 20 mM glucose. This diluted culture was added to the microtiter plate wells (100 μL/well) containing the peptide spots and controls. After 30 min of incubation, serial dilutions were performed from the membrane spots to successive rows of the plate. Luminescence of the *P. aeruginosa* PAO1 strain H1001 containing luciferase gene cassette *luxCDABE* was measured at 4 h using a Tecan Spectra Fluor Plus (Tecan US). Peptides were assayed in dilution series in sets of 10 peptides with one control peptide Bac2A per series.

Calculation of Peptide Activity. The luminescence of each peptide in a dilution series was fit to the following function (1) independently for each peptide, after luminescence data were normalized to 1.0 for the most dilute luminescence point for each peptide. This function had the form of a sigmoid curve consisting of two plateaus with a smoothly varying region joining them. Parameters of the function described the height of the plateau, the position of the center of the slope at half the maximum luminescence, and the slope at the center. Estimation of parameters was performed using custom C software using Numerical Recipes in C.[27]

$$L = \frac{L_{\max}}{1 + e^{-2S(x - x_{1/2})}} \quad (1)$$

In this function, $L_{\max}$ controlled the maximum height of the curve, S controlled the slope, and $x_{1/2}$ was the value of x giving luminescence of half of the maximum luminescence. The values of x were in dilution steps with values from 0, for the initial concentration, to 7 (after seven dilutions); these corresponded to changes in concentration C

$$C = C_0 2^{-x} \quad (2)$$

where C_0 was the initial concentration of peptide in the undiluted well. We were interested in calculating the concentration of peptide that reduced the number (and hence the luminescence) of viable energized bacteria by 50%, the IC_{50}. From these equations we could state the IC_{50} as

$$IC_{50} = C(x_{1/2}) = C_0 2^{-x_{1/2}} \quad (3)$$

However, we could eliminate the need to determine the initial concentration of peptide by reporting the activity of peptides as rel IC_{50} values: the ratio of IC_{50} for the experimental peptide to the IC_{50} for Bac2A. Values of rel $IC_{50} < 1.0$ mean the peptide is more active than Bac2A because a lower concentration yields the same reduction in bacterial numbers. For peptides with very low or zero activity, curve fitting was problematic. When the luminescence of a well for an undiluted peptide was >50% of the maximum luminescence for the peptide at high dilutions, the IC_{50} concentration was not observed even at the highest peptide concentration used. Here, the peptide was considered to be inactive and assigned a rel IC_{50} value of 25 (the approximate lower limit of activity that could be measured).

For sets A and B, seven dilution points were used in the calculation of rel IC_{50} due to frequent artifacts in the last dilution row (changes in luminescence were observed that were inconsistent with the expected profile). For the 200 peptides taken from the independent test set, the rel IC_{50} was determined from all eight dilution points for each peptide, because these artifacts were largely eliminated in later measurements.

QSAR Descriptors. The QSAR descriptors used in this study are shown in Supporting Information Supplementary Table 1. The "inductive" QSAR descriptors used in this study were previously described.[17] An initial set of 77 QSAR descriptors was calculated for each peptide in the two training and test sets using MOE (Molecular Operating Environment, 2005, Chemical Computing Group Inc., Montreal, Canada). For purposes of model training and screening, the peptide structure was optimized on the basis of an initial linear structure followed by potential energy minimization of each molecule using MMFF94 force-field calculations[28] with structure optimization done without including interactions with other molecules. [For determining model sensitivity to peptide structural conformation, we also performed prediction of set A peptide activities using MMFF94 force fields with peptide structures in (1) initial (straight) peptide backbone or structures resulting from energy minimization (2) in gas or (3) using a Born implicit solvation model. Comparison of predictions is in Supporting Information Supplementary Table 2.]

The atomic types have been assigned according to their name, valence state, and formal charge of constituent atoms, as defined within MOE. QSAR descriptors were calculated using custom SVL scripts within the MOE environment. The inductive QSAR variables can be computed by the following equations

$$R_{sj\rightarrow G} = R_j^2 \sum_{i\neq j}^{N-1} \frac{1}{r_{j-i}^2} \tag{4}$$

$$R_{sG\rightarrow j} = \alpha \sum_{i\subset G, i\neq j}^{n} \frac{R_i^2}{r_{i-j}^2} \tag{5}$$

$$\sigma^*_{j\rightarrow G} = \sum_{i\neq j}^{N-1} \frac{(\chi_j^0 - \chi_i^0)R_j^2}{r_{j-i}^2} \tag{6}$$

$$\sigma^*_{G\rightarrow j} = \beta \sum_{i\subset G, i\neq j}^{n} \frac{(\chi_i^0 - \chi_j^0)R_i^2}{r_{i-j}^2} \tag{7}$$

$$\chi^0_{G\rightarrow j} = \frac{\sum_{i\neq j}^{N-1} \frac{\chi_i^0(R_i^2 + R_j^2)}{r_{i-j}^2}}{\sum_{i\neq j}^{N-1} \frac{R_i^2 + R_j^2}{r_{i-j}^2}} \tag{8}$$

$$\Delta N_j = Q_j + \gamma \sum_{i\neq j}^{N-1} \frac{(\chi_j - \chi_i)(R_j^2 + R_i^2)}{r_{j-i}^2} \tag{9}$$

$$\eta_j = \frac{1}{2\sum_{j\neq i}^{N-1} \frac{R_j^2 + R_i^2}{r_{j-i}^2}} \tag{10}$$

$$\eta_{\mathrm{MOL}} = \frac{1}{s_{\mathrm{MOL}}} = \frac{1}{2\sum_{j\neq i}^{N-1} \frac{R_j^2 + R_i^2}{r_{j-i}^2}} \tag{11}$$

$$s_i = 2\sum_{j\neq i}^{N-1} \frac{R_j^2 + R_i^2}{r_{j-i}^2} \tag{12}$$

$$s_{\mathrm{MOL}} = \sum_{j\neq i}^{N} \sum_{j\neq i}^{N} \frac{R_j^2 + R_i^2}{r_{j-i}^2} \tag{13}$$

where R = covalent atomic radius, r = interatomic distance, Q_j = formal charge of atom j, χ = inductive electronegativity, R_S = steric constant, σ^* = inductive constant, ΔN = inductive partial charge, and η and s = inductive analogues of chemical hardness and softness, respectively.

It should be noted that the variables indexed with j subscript describe the influence of a single atom on a group of atoms G (typically the rest of the N-atoms molecule), whereas G indices designate group (molecular) quantities. The linear character of these equations makes inductive descriptors readily computable and suitable for sizable databases and positions them as appropriate parameters for large-scale QSAR models. Resources using the R language for statistical computing (http://www.r-project.org)[29] were used for all following steps. Each descriptor in the training and test sets was normalized to the range encountered in training peptide sets A and B. A cross-correlation (Pearson) was performed on the values of descriptors in the set of all peptides. The descriptors were ordered according to a priority number (indicated in Supporting Information Supplementary Table 1) to prioritize descriptors. When a descriptor correlated with one or more other descriptors at >0.95 or <−0.95, the descriptor with the lowest priority number was retained and the others were dropped from use in modeling. This left a final set of 44 descriptors (Supporting Information Supplementary Table 1).

Hydrophobic moments were calculated for comparison purposes (Table 2; Supporting Information Supplementary Table 5) and not used in ANN modeling. These were calculated using the *hmoment* utility in EMBOSS[30] modified to utilize the Eisenberg scale.[31]

Training and Validation Data Sets. For each of the three training sets of peptides described above (set A, set B, and set A+B), the peptides were classified by considering the top 5% of rel IC_{50} values to be active peptides and assigned the activity value of 1 in the data sets for training the ANNs; other activity values were assigned 0. A stratified 10-fold cross-validation was performed on the three sets, resulting in 10 models for each set for a total of 30 models. Briefly, to create the cross-validation data sets, 10% of the active peptides in the training set (one of set A, set B, or set A+B) were randomly assigned to each of 10 lists. Then 10% of the inactive peptides in the training set were randomly assigned to each of 10 lists. One list of actives was combined with one list of inactives, to create 10 lists of combined active and inactive peptides. Using one of these lists as the peptides for a validation data set, the other 9 were used as the corresponding training set. This was repeated a total of 10 times to create 10 validation sets and 10 training sets. This creation of 10-fold cross-validation sets was performed separately for each of the training sets (A, B, and A+B).

Test Data Set. To evaluate the voting system's ability to predict peptide activity, we selected a set of 100,000 peptide sequences according to the amino acid frequencies used in set B. QSAR descriptors were calculated as described above. The maximum and minimum values of each of the 44 descriptors were compared to the range present in the set A and B training data. When a peptide in the test data was outside 15% above or below the range in the training data, the test peptide was dropped from the test set, leaving a total of 99,577 peptide sequences.

Model Training. Exploratory data analysis was performed on set A data using PCR and PLSR to identify significant variables. Analysis was performed using the pls package[32] from the R Project (http://www.r-project.org).[29]

ANNs were constructed and evaluated using SNNS (Stuttgart Neural Network Simulator, version 4.2, from University of Tübingen, Stuttgart, Germany, available at http://www-ra.informatik.uni-tuebingen.de/SNNS/). The networks (Supporting Information Supplementary Figure 7) consisted of 44 input nodes (one for each QSAR descriptor as described above), 10 nodes in one hidden layer, and 1 output node; all were fully connected. The output node values for training were 0 for not active, and 1 for active. Networks were initialized using randomized weights.

Model training was performed using pairs of training and validation data sets generated for the 10-fold cross-validation described above. Therefore, 10 models were created for each of the training sets (set A, set B, and set A+B) for a total of 30 models. Training was performed on each training data set using the standard back-propagation learning function with parameters $\alpha = 0.2$ and $d_{\max} = 0$. The update function used topological order with shuffled order of training patterns. For each cycle of training, the validation data set was evaluated. As the network trained, network parameters giving a minimum error on the validation set were stored. After 200 training cycles with no new minimum model error found, all network weights were jogged by 2% to attempt to escape local minima; weights that show >95% correlation during propagation were jogged by 5%. Training continued and was terminated after an additional 200 cycles with no new minimum validation error encountered. Performance measures such as ROC curves and areas, sensitivity, and specificity were calculated using the ROCR package in R.[33]

In Silico Ranking and Selection of Test Peptides. To test the predictions of the ANNs, all peptides in the test set were evaluated by all 30 ANNs, and the combined predictions were integrated into a single ordering of the test peptides as follows. Each peptide in the test set was assigned a ranking by each ANN. If a test peptide appeared in the top 5% of all peptides in the test set for an ANN, it received one "vote" to indicate the model suggested it to be highly active. Therefore, a test peptide may receive up to 30 votes from the total of 30 ANNs. Peptides were ranked by number of votes with the relative ordering of peptides receiving the same number of votes determined by the average of the rankings of all ANNs.

Sets of 50 peptides at four positions of overall ranking were selected to independently evaluate the system's ability to predict peptide activity and inactivity. Quartile 1 (Q1) peptides were ranked in the topmost 50 positions and considered the most likely to be more active than control. Quartile 2 (Q2) peptides were ranked at the start of the second quartile, positions 24,895−24,944, and considered likely to be more active than control. Quartile 3 (Q3) peptides were ranked at the end of the third quartile, positions 74,673−74,682, and considered likely to be less active than control. Quartile 4 (Q4) peptides were ranked at the end of the fourth quartile, positions 99,568−99,577, and considered to be most likely less active than control. These 200 predicted peptides were synthesized and assayed for activity as described above.

Minimal Inhibitory Concentration (MIC) Determination. The MIC of the peptides was measured as described in ref 26. Briefly, a modified broth microdilution method was used. The peptides synthesized in bulk were dissolved and stored in glass vials. Peptide purity was assessed by HPLC and MS as shown in Supporting Information Supplementary Table 6 (15 of the 18 peptides have a purity of ≥95%).The assay was performed in sterile 96-well polypropylene microtiter plates (catalog no. 3790, Costar, Cambridge, MA). Serial dilutions of the peptides to be assayed were performed in 0.01% acetic acid containing 0.2% bovine serum albumin at 10-fold the desired final concentration. Ten microliters of the 10-fold concentrated peptides was added to each well of a 96-well polypropylene plate containing 90 μL of MH medium per well. Bacteria were added to the plate from an overnight culture at a final concentration of $(2-7) \times 10^5$ CFU/mL and incubated overnight at 37 °C. The MIC was taken as the concentration at which no growth was observed.

MIC analyses were done on a panel of bacterial pathogens that were both susceptible and resistant to common antibiotics. *P. aeruginosa* PAO1 strain H10319, *P. maltophilia* ATCC 13637, *S. aureus* ATCC 2592319, *E. faecalis* ATCC 292129, and *E. cloacae* 218R, constitutively expressing class C chromosomal β-lactamase 31, were from our laboratory strain collection. A methicillin -resistant *S. aureus* (MRSA) clinical isolate was kindly provided by Anthony Chow (Vancouver General Hospital, Vancouver, Canada). Two *K. pneumoniae* and two *E. coli* clinical isolates expressing extended spectrum β-lactamases (ESBL) were kindly provided by George Zhanel (Health Sciences Centre, Winnipeg, Canada). Vancomycin-resistant clinical isolates of *E. faecalis* and *E. faecium* were obtained from Ana M. Paccagnella (BC Centre for Disease Control, Vancouver, Canada). Three clinical isolates (9, 198, and 213) of multidrug-resistant *P. aeruginosa* were kindly provided by Carlos Kiffer (University of São Paulo, Brazil). These isolates all have resistance to piperacillin/tazobactam, Meropenem, ceftazidime, ciprofloxacin, and cefepime, and 9 is also polymyxin B resistant. Three *P. aeruginosa* clinical isolates of the Liverpool epidemic strain (LES) (H1027, H1030, and LES400) 32 were all kindly provided by Craig Winstanley (University of Liverpool, U.K.). LES400 was resistant to gentamicin and tobramicin, whereas H1030 showed resistance to colistin, amikacin, gentamicin, and tobramicin. All tested bacterial strains were categorized as biohazard level 2 pathogens.

Conclusions

We have demonstrated in this study the specific methodology used in the first application of atomic resolution 3D QSAR methodology prediction of antibacterial activity to a large data set of diverse peptides. With the availability of large numbers of synthetic peptides and a rapid assay to determine their antibacterial activity, larger sets of data on peptide sequence and activity can now be created. On the basis of two random libraries containing a total of >1400 peptides, we developed artificial network models that predict and rank the relative activities of novel antimicrobial peptides with remarkable accuracy: in an independent test set of 100,000 virtual peptides, 94% of the 50 highest ranked peptides predicted to be highly active were found to be highly active.

In addition to creating more complex models that utilize the inductive QSAR methodology, the availability of high-quantity and -quality peptide data also allows more rigorous training and evaluation of the machine learning techniques. We consider the methodology described here to be the first successful demonstration of high-throughput in silico screening of antibacterial peptides for novel drug leads.

Acknowledgement. We gratefully acknowledge financial support from the Canadian Institutes for Health Research (CIHR) and the Foundation of the National Institutes of Health and CIHR through the Grand Challenges in Global Health Initiative. We thank Jessica Lee for technical support in creating the computer-based peptide libraries. R.E.W.H. is the recipient of a Canada Research Chair. K.H. received a CIHR fellowship. C.D.F. received a Doctoral Research Award from the CIHR.

Supporting Information Available: Supplementary Table 1, inductive and conventional molecular descriptors for the QSAR modeling; Supplementary Table 2, effect of conformation on ANN prediction results; Supplementary Table 3, correlation between descriptors used in analysis of antibacterial activity; Supplementary Table 4, PLSR loadings for first six components; Supplementary Table 5, candidate peptides for confirmation of QSAR predictions; Supplementary Table 6, peptide purity; Supplementary Figure 1, occurrence of amino acids in the training and QSAR predicted data sets; Supplementary Figure 2, fraction of activity explained by PLSR; Supplementary Figure 3, fraction of activity explained by PCR; Supplementary Figure 4, predicted activity using PLSR; Supplementary Figure 5, loadings on components of PLSR model; Supplementary Figure 6, impact of ANN settings on performance; Supplementary Figure 7, structure of an artificial neural network; Supplementary Figure 8, receiver operating characteristic curves for the three data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Finlay, B. B.; Hancock, R. E. W. Can innate immunity be enhanced to treat microbial infections? *Nat. Rev. Microbiol.* **2004**, *2*, 497–504.
(2) Hamilton-Miller, J. M. T. Antibiotic resistance from two perspectives: man and microbe. *Int. J. Antimicrob. Agents* **2004**, *23*, 209–212.
(3) Hancock, R. E. W.; Sahl, H. G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat. Biotechnol.* **2006**, *24*, 1551–1557.
(4) Koczulla, A. R.; Bals, R. Antimicrobial peptides: current status and therapeutic potential. *Drugs* **2003**, *63* (4), 389–407.
(5) Levy, S. B.; Marshall, B. Antibacterial resistance worldwide: causes, challenges and responses. *Nat. Med.* **2004**, *10*, S122−S129.
(6) Jenssen, H.; Hamill, P.; Hancock, R. E. W. Peptide antimicrobial agents. *Clin. Microbiol. Rev.* **2006**, *19* (3), 491–511.
(7) Perkins, R.; Fang, H.; Tong, W.; Welsh, W. J. Quantitative structure−activity relationship methods: perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.* **2003**, *22*, 1666–1679.
(8) Jenssen, H.; Gutteberg, T. J.; Lejon, T. Modeling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. *J. Pept. Sci.* **2005**, *11*, 97–103.
(9) Lejon, T.; Stiberg, T.; Strom, M. B.; Svendsen, J. S. Prediction of antibiotic activity and synthesis of new pentadecapeptides based on lactoferricins. *J. Pept. Sci.* **2004**, *10*, 329–335.
(10) Lejon, T.; Strom, M. B.; Svendsen, J. S. Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J. Pept. Sci.* **2001**, *7*, 74–81.
(11) Strom, M. B.; Stensen, W.; Svendsen, J. S.; Rekdal, O. Increased antibacterial activity of 15-residue murine lactoferricin derivatives. *J. Pept. Res.* **2001**, *57*, 127–139.
(12) Frecer, V. QSAR analysis of antimicrobial and haemolytic effects of cyclic cationic antimicrobial peptides derived from protegrin-1. *Bioorg. Med. Chem.* **2006**, *14*, 6065–6074.
(13) Frecer, V.; Ho, B.; Ding, J. L. De novo design of potent antimicrobial peptides. *Antimicrob. Agents Chemother.* **2004**, *48*, 3349–3357.
(14) Ostberg, N.; Kaznessis, Y. Protegrin structure−activity relationships: using homology models of synthetic sequences to determine structural characteristics important for activity. *Peptides* **2004**, *26*, 197–206.

(15) Jenssen, H.; Fjell, C. D.; Cherkasov, A.; Hancock, R. E. QSAR modeling and computer-aided design of antimicrobial peptides. *J. Pept. Sci.* **2008**, *14* (1), 110–114.
(16) Jenssen, H.; Lejon, T.; Hilpert, K.; Fjell, C. D.; Cherkasov, A.; Hancock, R. E. Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward *P. aeruginosa*. *Chem. Biol. Drug Des.* **2007**, *70* (2), 134–142.
(17) Cherkasov, A. 'Inductive' descriptors. 10 successful years in QSAR. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 21–42.
(18) Cherkasov, A. Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. *Int. J. Mol. Sci.* **2005**, *6*, 63–86.
(19) Karakoc, E.; Cherkasov, A.; Sahinalp, S. C. Distance based algorithms for small biomolecule classification and structural similarity search. *Bioinformatics* **2006**, *15*, 243–251.
(20) Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.
(21) Hilpert, K.; Volkmer-Engert, R.; Walter, T.; Hancock, R. E. W. High-throughput generation of small antibacterial peptides with improved activity. *Nat. Biotechnol.* **2005**, *23*, 1008–1012.
(22) Cherkasov, A.; Hilpert, K.; Jenssen, H.; Fjell, C. D.; Waldbrook, M.; Mullaly, S. C.; Volkmer, R.; Hancock, R. E. W. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chem. Biol.* **2009**, *4* (1), 65–74.
(23) Sawyer, J. G.; Martin, N. L.; Hancock, R. E. Interaction of macrophage cationic proteins with the outer membrane of *Pseudomonas aeruginosa*. *Infect. Immun.* **1988**, *56* (3), 693–698.
(24) Hilpert, K.; Elliott, M. R.; Volkmer-Engert, R.; Henklein, P.; Donini, O.; Zhou, Q.; Winkler, D. F.; Hancock, R. E. Sequence requirements and an optimization strategy for short antimicrobial peptides. *Chem. Biol.* **2006**, *13* (10), 1101–1107.
(25) Yeaman, M. R.; Yount, N. Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* **2003**, *55*, 27–55.
(26) Hilpert, K.; Winkler, D. F.; Hancock, R. E. Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. *Nat. Protoc.* **2007**, *2* (6), 1333–1349.
(27) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannerty, B. P. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: New York, 1992.
(28) Halgren, T. A. Merck molecular force field 0.1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490–519.
(29) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.
(30) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277.
(31) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81* (1), 140–144.
(32) Mevik, B. H.; Wehrens, R. The pls package: principal component and partial least squares regression in R *J. Stat. Software* **2007**, *18*, (2).
(33) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **2005**, *21*, 3940–3941.

JM8015365