Chem Biol Drug Des 2011; **77:** 48–56

Research Article

doi: 10.1111/j.1747-0285.2010.01044.x

Optimization of Antibacterial Peptides by Genetic Algorithms and Cheminformatics

Christopher D. Fjell[1,2,*], Håvard Jenssen[2], Warren A. Cheung[1], Robert E. W. Hancock[2] and Artem Cherkasov[3]

[1]Faculty of Medicine, Division of Infectious Diseases, Department of Medicine, University of British Columbia, 2733 Heather Street, Vancouver, BC, V5Z 3J5, Canada
[2]Centre for Microbial Diseases and Immunity Research, University of British Columbia, 2259 Lower Mall, Vancouver, BC, V6T 1Z4, Canada
[3]Prostate Centre at the Vancouver General Hospital, University of British Columbia, 2640 Oak Street, BC, V6H 3Z6, Canada
**Corresponding author: Christopher D. Fjell, cfjell@interchange.ubc.ca*

Pathogens resistant to available drug therapies are a pressing global health problem. Short, cationic peptides represent a novel class of agents that have lower rates of drug resistance than derivatives of current antibiotics. Previously, we created a software system utilizing artificial neural networks that were trained on quantitative structure-activity relationship descriptors calculated for a total of 1400 synthetic peptides for which antibacterial activity was determined. Using the trained system, we correctly identified additional peptides with activity of 94% accuracy; active peptides were 47 of the top rated 50 peptides chosen from an *in silico* library of nearly 100 000 sequences. Here, we report a method of generating candidate peptide sequences using the heuristic evolutionary programming method of genetic algorithms (GA), which provided a large (19-fold) improvement in identification of novel antibacterial peptides. Approximately 0.50% of peptides evaluated during the GA method were classified as highly active, while only 0.026% of the nearly 100 000 sequences we previously screened were classified as highly active. A selection of these peptides was tested *in vitro* and activities reported here. While GA significantly improves the possibility of identifying candidate peptides, we encountered important pitfalls to this method that should be considered when using GA.

Key words: antibacterial peptides, cheminformatics, genetic algorithms, quantitative structure-activity relationship

Abbreviations: AMP, antimicrobial peptide; ANN, artificial neural network; HPLC, high-performance liquid chromatography; MIC, minimum inhibitory concentration; MRSA, methicillin-resistant *Staphylococcus aureus*; MS, mass spectroscopy; QSAR, quantitative structure-activity relationship; Rel.IC_{50}, relative inhibitory concentration 50% (relative to control peptide Bac2A); VRE, vancomycin-resistant *Enterococcus*.

Received 10 June 2010, revised 15 September 2010 and accepted for publication 19 September 2010

Human pathogens that are resistant to current antibiotic treatments represent a significant health threat worldwide (1). Drugs based on synthetic peptides are inspired by the short cationic, amphipathic peptides found throughout the kingdoms of life that possess antimicrobial activity by various mechanisms (2). These peptides have drawn significant attention as a possible source of novel antibacterial agents (3–6). While antimicrobial peptides generally exhibit lower potency against susceptible bacterial targets compared to conventional low-molecular-weight antibiotic compounds, they have advantages that compensate for this lower potency, including fast killing, a broad range of activity, a postulated multiplicity of targets, low toxicity for host cells, effectiveness against clinically multidrug-resistant bacteria, and minimal development of resistance in target organisms (6,7).

We have recently shown that synthetic peptides with high antibacterial activity and low toxicity can be identified with high accuracy using cheminformatics and machine learning and without the use of an original template sequence (8,9). To achieve this, we used a quantitative structure-activity relationship (QSAR) approach utilizing artificial neural networks (ANN) to build computational models of peptide activity based on data from over 1400 random sequences, biased to contain amino acids believed from substitution analyses to be important for antibacterial activity. As a basis for describing structure in these peptides, a set of 44 descriptors were employed, including 3D QSAR descriptors that utilize atomic-scale molecular information, the so-named 'inductive' QSAR descriptors reviewed previously (10). Briefly, the 'inductive' QSAR descriptors describe whole molecules based on the calculated effects of the atomic constituents of a molecule. A total of 26 'inductive' descriptors were used in this study (Table S1), based on electronegativity, hardness, charge, substituent, and steric effects. An additional 18 conventional whole-molecule descriptors were used, including numbers of hydrogen acceptors and donors, surface area, total and partial charges, and molecular weight. In addition to peptide studies, these have been successfully applied to a number of molecular modeling studies, including identification of antibacterial activity of small compounds (11) and classification of antimicrobial compounds, conventional drugs, and drug-like substances (12,13).

To demonstrate the effectiveness of these techniques in identifying drug candidates, an *in silico* screening of 100 000 peptides was performed. By randomly synthesizing example peptides from each activity quartile, including low-activity peptides, it could be demonstrated that peptides with superior activity could be identified with 94% accuracy. However, the complexity of the ANN solution prevented us from 'inverting' the solution and using it to directly determine peptide sequences that are predicted to be active; instead, a small number of active peptides were identified from a large set of *in silico* candidates by computational evaluation.

An exhaustive search was not possible because of the large number of possible peptide variants (X^{20}, where X is the number of amino acids in the peptide chain) and the time and resources needed for QSAR descriptor calculations. Thus, it is advantageous to utilize a search strategy that minimizes the number of peptides that need to be evaluated to determine additional highly active peptides. Here, genetic algorithms (GA) was applied to this problem as these evolutionary methods have been applied successfully in other areas of cheminformatics (14–17). [NB. an earlier application to a modest number of antimicrobial peptides was reported in the context of patents (18,19).]

A genetic algorithm is a heuristic method for search-and-approximation problems and is particularly well suited for problems involving string-like data such as the amino acids in a peptide. GA operates on populations of solutions by iteratively enhancing solutions using operations inspired by natural genetic processes: recombination (combining parts of two solutions to suggest another) and mutations (randomly changing one part of a solution to generate another). Each solution is composed of elements that are randomly modified ('mutated') or shuffled with other solutions ('recombined') and evaluated for fitness at each iteration ('generation'). The best solutions are propagated into the next iteration with new solutions added to the population based on modifications and combinations of these best peptides. A genetic algorithm solution requires that the problem be described in terms of a genetic representation with a fitness function specified to evaluate each solution. The genetic algorithm then either passes high-fitness individuals on to the next generation, removes low-fitness individuals, or creates offspring by recombination of two existing individuals or by mutation of an existing individual. Examples of mutation and recombination that showed dramatic changes on peptide fitness are shown in Figure 1, whereby mutation of one amino acid (V–I) increased fitness (described later) from 20 to 26, while recombination of two peptides with fitness 20 yielded a peptide with fitness 0.

In our previous studies, we created a software system to predict the rank the activity of 9-mer peptides, producing a fitness score. This system was constructed to make maximal use of the available experimental data by utilizing models produced by a stratified 10-fold cross-validation, as described previously (8,9). The system consisted of a set of 30 ANN models derived from the 10-fold cross-validation models of two datasets (Set A and B) of screened peptides plus the combined set (Set A + B). These were classification models trained to consider the top 5% as active. Confidence that any given peptide was active could be judged by the number of models that classified the peptide as active. As reported previously, the accuracy of prediction of peptide activity was highest when the largest numbers of models predicted activity: for example, for the top 50 peptides predicted of a set of 100 000 amino acid-biased semi-random peptides, the number of models indicating high activity ranged from 25 to 29 of 30. For these peptides, the accuracy of predicting highly active peptides was 94%. This number of models indicating high activity was therefore taken as the genetic algorithm fitness score.

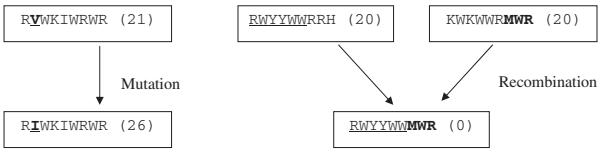

Figure 1: Examples of peptide evolution. Two examples of peptide evolution are shown: mutation of a single amino acid that results in an improved peptide and recombination of two moderate scoring peptides recombining to form one low-scoring peptide. Values in round brackets are the fitness scores for the peptides.

In the current work, we demonstrate that a genetic algorithm approach dramatically lowers the number of peptides that must be evaluated for *in silico* screening of synthetic peptides to identify those with high potency. In our previous work, we produced software models based on molecular descriptors that reliably predicted peptide antibacterial activity. However, in the previous work, peptides sequences were drawn from a large population of sequences based on assigned probabilities for each amino acid (without restriction within a peptide). Here, we remove the restriction of specifying the amino acid frequency and use GA to produce novel peptide sequences. While dramatically reducing the computational efficiency, we demonstrate that GA solutions are dependent on the starting population with dramatically differing results based on initial peptide population.

Materials and Methods

Peptide fitness function and classification models for highly active peptides

As described previously (8,9), we had constructed a software model to classify peptides as highly active or inactive based on a set of 44 QSAR descriptors calculated for each peptide combined with ANNs (see Figure S4 for diagram) trained on measured activity of 1433 peptides. The system consisted of 30 ANNs, each classifying a peptide as in the top 5% of activity of peptides with measured activity. Briefly, each ANN was configured with three layers: one input layer consisting of 44 input nodes (one per descriptor), one hidden layer of 10 hidden nodes, and one output layer with one node. Each layer was fully connected to the adjacent layer. We had measured the activity of 1433 peptides in two sets (of 933 and 500 peptides) using a luminescence assay, where killing is implied by a decrease in constitutive luminescence in *Pseudomonas aeruginosa* strain H1001 containing luciferase gene cassette *luxCDABE*. We used a stratified 10-fold cross-validation for the two sets of peptides and the sets combined (for a total of three sets). ANNs were trained to predict whether each peptide was in the top

5% of activity of the peptides in the set using a standard back-propagation so that internal connections (weights) between the input and hidden layers, and the hidden and output layers were optimized to reduce the error between measured and predicted activity at the output node. In the present work, we used the same threshold values to classify novel peptides using the number of models predicting high activity as the fitness score for the genetic algorithm. Additional details are in Supporting Information.

Initial peptide population

Five simulated evolution experiments were performed here. Two small initial populations of peptides were selected from the population of random 100 000 peptides used previously (8,9). (The amino acid frequency of the population is shown in Table S3. No restriction was placed on the composition of amino acids in each peptide for this original population; these are simply the probability of choosing an amino acid for any position in any peptide.) Peptides were chosen for initial populations in the GA simulations to maximize the diversity of amino acids present in the population. Peptides containing all the 12 amino acids (F, G, H, I, K, L, M, Q, S, T, V, and Y) present in the population were selected at two levels of fitness score. We chose a small starting number of peptides for three reasons: (i) to reduce the cost of peptides were to be synthesized, (ii) to rapidly evaluate the GA method by keeping the population of peptides small and increase the number of simulated generations, and (iii) to balance the amino acid types. Of the 12 amino acids types found in the population of moderate activity peptides (fitness score of 20 or 21), three (G, Q, S) were found in only one peptide. Therefore, for these reasons, we chose to take only two peptides representing each of the 12 types. We randomly selected two peptides for each amino acid type then removed five peptides to further reduce the size (because a single peptide contains more than one desired amino acid type). In simulation A, 19 peptides with moderate activity were selected from 100 000 peptides biased random population having moderate prediction of activity (fitness score of 20 or 21). Similarly, in simulation B, 22 peptides were selected having a fitness score of 2. Further simulations (C and D) were run from initial populations of peptides having equal probability of all amino acids except cysteine (which was excluded). Simulation E was run from a starting population containing only the five amino acids of highest composition in the experimental peptides (I, K, R, V, W). Randomization was performed using runif() function of R specifying the appropriate amino acid array.

Evolution of peptide sequences

The initial populations of peptides were evolved over 600 generations using custom Java code utilizing the JGAP 3.2 (http://jgap.sourceforge.net) genetic algorithm package and converting single letter amino acid peptides into integer arrays for manipulations. QSAR descriptors were calculated through embedded calls to Molecular Operating Environment 2007.09 (MOE, 2007; Chemical Computing Group Inc., Montreal, QC, Canada) from the Java code. 'Inductive' descriptors were calculated using custom scientific vector language (SVL) scripts using MOE, while conventional descriptors were calculated using standard MOE methods. The population size was allowed to vary to ensure all high-scoring peptides remained in the population. A mutation rate of one change per 15 amino acids was used.

Assessment of predictor domain

Leverage points were calculated to assess whether peptides occupied the same domain, as described in (20) using the R statistical language. Leverage points were calculated for linear models relating 1433 peptides descriptors to the IC50 value and 8249 peptides from simulations to the fitness values of the peptides. As discussed in (21), values of leverage points were compared to the 'warning leverage', h^*, value of $3k/n$, where k is the number of parameters in the linear model (=45) and n is the number of training samples (n = 1433, h^* = 0.0942 for experimental IC50 peptides; and n = 8,249, h^* = 0.0164 for simulated peptides).

Evaluation of peptide antibacterial activity

Antibacterial activity of synthesized peptides was determined [as previously described (9,22)] using a luminescence-based *in vitro* assay and reported as inhibitory concentration at 50% relative to a control peptide (Rel.IC_{50}) using Bac2A as the control peptide. Briefly, a culture of *P. aeruginosa* PAO1 strain H1001 (containing a luciferase gene cassette *luxCDABE*) was treated with peptide at eight concentrations. The concentration of peptide required to reduce luminescence by 50% was estimated by curve fitting using custom software.

Minimal inhibitory concentration (MIC) determination

The MIC of the peptides was measured as previously described (8,9). 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 high-performance liquid chromatography as shown in Table S2 (12 of the 14 peptides had a purity of 95% or greater). The MIC assay was performed in sterile 96-well polypropylene microtitre plates (Cat. #3790; Costar, Costar, Cambridge, MA, USA). 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 microliter of the 10-fold concentrated peptide stocks were added to each well of a 96-well polypropylene plate containing 90 μL of Mueller-Hinton (MH) media 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.

The following microbes were tested for MIC: *P. aeruginosa* PAO1 strain H103 (23), *Staphylococcus aureus* ATCC#25923 (23). An methicillin-resistant *S. aureus* clinical isolate was kindly provided by Anthony Chow (Vancouver General Hospital, Vancouver, BC, Canada). Vancomycin-resistant clinical isolates of *Enterococcus faecium* were obtained from Ana M. Paccagnella (BC Centre for Disease Control, Vancouver, Canada). A clinical isolate of *Escherichia coli* expressing extended spectrum β-lactamases (ESBL) were kindly provided by George Zhanel (Health Sciences Centre, Winnipeg, Canada). A clinical isolate (#213) of multidrug-resistant *P. aeruginosa*

was kindly provided by Carlos Kiffer (University of São Paulo, Brazil). These isolates all have resistance to piperacillin/tazobactam, meropenem, ceftazidime, ciprofloxacin, and cefepime. *P. aeruginosa* clinical isolates of the Liverpool epidemic strain (LES400) (24) were all kindly provided by Craig Winstanley (University of Liverpool, UK). LES400 was resistant to gentamicin and tobramicin. *Candida albicans* were obtained from Barbara Dill (UBC department microbiology, Vancouver, Canada). All tested bacterial strains were categorized as biohazard level 2 pathogens.

Results and Discussion

Evaluation of peptide fitness score

As described previously, peptide fitness score was taken as the number of ANNs predicting the peptide as active. As our software system uses a consensus of 30 ANNs, the fitness score varies from 0 to 30.

Initial population of peptides

Genetic algorithm searches were executed starting from five initial populations of peptides. There were several goals: first, to identify additional peptides with very high-fitness scores to evaluate the ability of GA to identify novel peptides for screening by antibacterial activity assay and secondly, to understand the importance of starting population on the composition of later peptide populations in a search. For the first two simulations (A and B), both populations of peptides were selected from the set of 100 000 peptides described previously (8,9) at different levels of fitness score. Briefly, this set of 100 000 sequences had a specified probability for each amino acid, regardless of location or other amino acids per peptide. For the first search (simulation A), peptides were selected that were moderately predicted to be active, having a fitness value of 20 of 30. However, the amino acid diversity was low for those peptides, and the fitness range was expanded to include those with fitness of 21. A small initial population of 19 peptides were selected to maximize the diversity of amino acids present in peptides with these initial fitness scores, ensuring that all amino acids present in the library were present to at least some degree in these peptides (Table 1), described later. Thus, the initial set of 19 peptides included all 12 amino acids present in the 594 peptides of the 100 000 peptide set having a fitness value of 20 or 21. As some amino acids had low representation (only one peptide containing any of G, Q and S, and two for H), it was decided to use a small population to minimize the effects of the relatively large numbers of certain other amino acids in the population. Similarly, the initial peptides for simulation B were selected from the total of 2503 peptides of the 100 000 having a fitness score of 2, a low score indicating low confidence that these are highly active peptides (Table 2). A random population of 22 peptides was selected and represented all amino acid types in this population in at least two peptides.

Three additional simulations were run for 200 generations each (because of time constraints for analysis) to assess the global evolution of the GA process. Two simulations (C and D) were run starting from 20 initial random peptides have equal probability of each amino acid at any position (except for cysteine, which was not included because of incompatibility with the synthesis method used previously for experimental peptides). One final simulation (E) was run from a population of 20 peptides randomized for only five amino acids (I, K, R, V, W) that are highly represented in the origi-

Table 1: Initial peptide population for simulation A. Peptides were chosen from a set of biased random sequences that had a score of 20 or 21 in simulation A (moderate confidence in activity) and selected to have diverse amino acids populations

Sequence	Score
KKWWYWWKR	20
KWKRWFKWR	21
KWKWWRMWR	20
MWRKWRRWW	21
RKKWWWLFR	21
RLKWWRWRW	21
RRWRWWWVW	21
RRWWWRLWW	21
RRWWWRRWY	21
RVWKIWRWR	21
RWIRKIWWR	21
RWIWWRRWW	21
RWRWWGWRR	20
RWRWWWKKT	20
RWWRWWKQR	20
RWWWWSRRR	20
RWYYWWRRH	20
RYRWWKWRH	20
TWWWKKWRR	20

Table 2: Initial peptide population for simulation B. Peptides were chosen from a set of biased random sequences that had a fitness score of 2 (low confidence in activity). Peptides were selected to have diverse amino acids populations

Sequence	Score
ARKWWWRWK	2
AWWRKRKWW	2
FVKRWWRFR	2
IGWWWRKRW	2
IWKRWWRKT	2
KNWKWWRWR	2
KRRSWWKWW	2
KRWRWLRWG	2
KWWRWRRFI	2
QRRRWWWWK	2
RLIRWWIRK	2
RRKRLYWIW	2
RRRWYWKWN	2
RRWRIWWIK	2
RTYKRWYRW	2
RWIRWWRQW	2
RWRHIWWRW	2
RWWKWRWLM	2
RWYKHWRFR	2
SRWWKRRWY	2
VKRWWWRRM	2
WWRKLWRKL	2

nal population of experimental peptides (above 5% of total; see Table S3). Both simulations C and D failed to find any peptide with fitness score above 10 but did converge on higher scoring subsequences (eg. starting sequence SDD for simulation C; EKWW and LLWW for simulation D). Simulation E initially contained a high-scoring peptide sequence IWWRWWIRR and converged on high-scoring subsequences IWKRW and KRWR. (See Tables S4 and S5).

Iterative improvement in peptides

As shown in Figure S1, there was rapid improvement in scores from the first generation to generation 100 with continued improvement up to generation 600 for simulations A and B. As expected, throughout the evolution of the population of peptides, the genetic algorithm created a set of peptides having a variety of fitness scores owing to the random nature of novel peptide generation. For simulation A, the final generation contained 34 peptides, including 10 peptides with score of 29 and 22 peptides with scores that were 26 or higher (Table 3). The highest score observed in any of the peptides studied here or previously (9) was 29 rather than the maximum of 30 possible. This suggests that the genetic algorithm method could not identify any peptides with a higher score than those already identified. Of the 10 top-scoring peptides, nine were closely related and started with the sequence RWKRW. There were three other peptides starting with this sequence with modestly to significantly lower scores, 28 (RWKRWWRIL), 21 (RWKRWWKVW), and 1 (RWKRWSRLL). The population of peptides always contained a proportion of lower scoring peptides (as seen in Figure S1) owing to the random nature of creation of novel peptides by the genetic algorithm. As well, there was a rapid increase in peptide fitness for simulation B, for which the initial population containing much lower scores as seen in Figure S1. Thus, the first generations showed a dramatic rise in fitness scores (Figure S2). The final peptide population for simulation B is shown in Table 4, with 25 of the 51 peptides scoring 26 or higher. The fitness scores for simulation A (mean 22.4, SEM 1.6) and simulation B (mean 20.9, SEM 1.2) were not significantly different (p-value >0.05 using Student's *t*-test). None of these peptides in the final populations were found in the previous 100 000 peptide population from our previous studies (8,9).

Table 3: Final peptide population simulation A. The final generation (generation 600) of peptides was sorted by score. The common subsequence RWKRW is shown in bold and discussed in the text

Sequence	Code	Fitness Score
RKRWWWRWW		29
RWKRWIRWW		29
RWKRWLRWW		29
RWKRWWRIW		29
RWKRWWRLL	GN-1	29
RWKRWWRLW		29
RWKRWWRVW		29
RWKRWWRWI	GN-2	29
RWKRWWRWL		29
RWKRWWRWW		29
KKRWWWWFR		28
KRWWWWKFR		28
KWWRWRRWW	GN-3	28
RKRWWWRWL		28
RWKKWWRWL	GN-4	28
RWKKWWRWW		28
RWKRWWRIL		28
KKRWWWWWR	GN-5	27
KWKRWRRWW		27
KWKRWWWWR		27
RKRWWWWFR	GN-6	27
KWKRWWWFR	GN-7	26
RKRWWWRWR		22
RWKRWWKVW		21
RWKWWWKFR		20
RWKKWWRVW		19
RWYRWWRIW		15
KRWRWWRLL		12
KWKKWWRWL		9
KWKRWWWWL		9
KKKRWRRWW		8
RWKYWWRII		4
RKRWWWRGL		1
RWKRWSRLL		1

There were two peptides (KWKRWWWFR and KWKRWWWWR) in common between the final populations for simulation A and B. Simulation B had no peptides with fitness score above 28 but included more peptides with high score (25 peptides with fitness scores of 26 and above). This suggests that the specific peptides in the final population were largely dependent on the initial population of peptides, as expected because the dominant method of generation of novel sequences was through cross-over from previous peptides and the effects of mutations were minor given the genetic algorithm parameters used here. The number of high-fitness score peptides appeared to be unchanged between generation 400 and generation 600 for both simulations A and B (Figure S1), suggesting that in each case, the genetic algorithm had settled on a 'local optimum' set of sequences from which it was unlikely to escape through continued evolution. Further improvements would likely require introduction of peptides with dramatically different sequences into the population.

Evolution of amino acid composition

The amino acid distribution of the peptide populations varied during the peptide sequence evolution (Figure S3). As described previously, the number of amino acid types was maximized when selecting the initial population to include 14 amino acid types for simulation A and 16 amino acids for simulation B. During evolution over the 600 generations, the number was reduced to seven amino acid types (in declining proportion: W, R, K, L, I, F, V) for the high-scoring peptides in simulation A and six amino acid types (in declining proportion: W, R, K, I, F, L) for the high-scoring peptides in simulation B. This proportion of amino acids for high-scoring peptides is similar to the proportions found previously for high-scoring peptide based on peptides sampled from a biased random library of 100 000 peptides (8,9).

Evaluation of peptide domain

In prediction studies, it is important to assess how similar a set of samples are to the samples used to generate the predictive model.

Table 4: Final peptide population, simulation B. The final generation (generation 600) of peptides is sorted by score. The common subsequence RWKRW is shown in bold and discussed in the text. Two peptides appear in both final populations (see also Table 3): KWKRWWWFR and KWKRWWWWR

Sequence	Code	Fitness Score
IWKRWWWKR	GN-8	27
KWKRWWWIR		27
KWKRWWWWR		27
RIWKIWWKR	GN-9	27
IKKRWWWFR	GN-10	26
IKWKRWWWR	GN-11	26
KLKRWWWFR		26
KLKRWWWWR		26
KWKRWWWFR		26
KWWKIWRWR	GN-12	26
KWWKRWKWR		26
KWWKRWWIR		26
KWWKRWWKR		26
KWWKRWWWR		26
RFWKIWWKR		26
RIWKRWWFR	GN-13	26
RLWKIWWRR		26
RLWKRWWFR		26
RLWKRWWIR	GN-14	26
RWWKIWKWR		26
RWWKIWWKR		26
RWWKIWWRR		26
RWWKRWWFR		26
RWWKRWWIR		26
RWWKRWWWR		26
IKKRWWWWR		25
KLKRWWWIR		25
KWWKIWWKR		25
KWWKRWWFR		25
RIWKRWWWR		25
RLKRWWWFR		25
RWKRWWWFR		25
KLWKRWWWR		24
RWWKIWRWR		24
KWWKIWKWR		22
RWWKWWWIR		22
RFWKIWRWR		21
KWKRIWWKR		19
RWWKRWAIR		19
RTWKRWWIR		18
RTWKIWKWR		12
KWWKRWWIH		11
KWWKRWSWR		10
RLWTRWWFR		9
RIWARWWFR		7
KWWKDWWKR		6
RFEKIWWKR		6
RIDKIWLKR		5
RLWKNWWRR		2
RFWQIWRWR		0
RWSKRWWWV		0

For example, the set of peptides with high activity had a strong bias in amino acid composition; the initial set of 1433 randomized peptides were also biased in amino acid composition regardless of activity. The question remains as to how similar the peptides evaluated during the GA evolution were to those used to train the models. The calculation of *leverage* is a standard method for assessing whether a sample is unusual in the space of the predictors (the space of chemical descriptor values in this case), and therefore less likely to be modeled well (20). Low-leverage values (near 0) indicate the sample is not unusual in the space of predictors, while large values indicate unusual points and possible problems in prediction. The warning leverage value, $h^* = 3k/n$, indicates where predictions for samples should be considered less reliable and depends on the number of variables in the model (k), as well as the number of samples (n) (21). To give an estimate of the confidence in prediction, we calculated leverage for the set of 1433 experimental peptides and those found from simulations C and D, which started from random initial peptides. We find that few peptides from the initial 1433 experimental peptides have leverage above the warning level (six of 1433, 0.42%; Table S5), while 43 of the 8249 (0.52%) peptides evaluated in the simulation from random peptides in simulation C and D gave leverage above warning leverage (Table S6). (Leverage is plotted against half-quantile values to illustrate the distribution of leverage compared to that expected from a normal distribution, Figure S5.)

Assessment of genetic algorithm performance

In a previous study, we examined 100 000 peptides from a random library of sequences that were biased with respect to frequency of amino acids. We empirically tested the activity of the 50 peptides ranked highest by fitness score. As we reported previously, 94% of these peptides were found to be highly active. This group of highly active peptides included all peptides with fitness scores of 26–29, and some peptides scoring 25 (some peptides scoring 25 were also outside of this group.). Therefore, to permit direct comparisons, it was considered here that peptides receiving a fitness score of 26 or higher could be relatively confidently predicted to have high antibacterial activity. A total of 22 peptides scoring ≥26 were previously identified (8,9) by computationally evaluating 99 576 peptides in the random library (=100 000 peptides−duplicates), or 0.026% highly active peptides. In contrast, using GA, we identified, over all generations of the simulated evolution of peptide populations, 22 peptides scoring ≥26 by evaluating a total of 4492 peptides (0.49% highly active) in simulation A and 25 peptides scoring ≥26 of 5067 peptides (0.51% highly active) in simulation B. Taking these two values as representative of the two methods (0.026% success when searching a large-biased random library using ANNs and 0.50% combined for the genetic algorithm search), a 19-fold enhancement in the discovery of highly active peptides was observed. As the progressive clustering of peptide scores into the high-scoring region was much slower after the first 100 generations, it seems likely that stopping the genetic algorithm at approximately generation 100 would be more efficient in terms of computational cost per highly active peptide found, because further highly active peptides will not be efficiently identified after this point. However, this increase in efficiency will prevent the algorithm from continuing to explore other possible peptide candidates and may reduce the diversity of peptides examined.

Minimal inhibitory concentrations were measured for fourteen selected peptides against a variety of pathogens, seven each from final populations of simulation A (codes GN-1 to -7) and simulation

Table 5: Activities against multiresistant microbes of selected peptides after simulated evolution. Values are μg/mL, measured in 3–6 repeated experiments

	Minimum inhibitory concentration (μg/mL)													
	GN-1	GN-2	GN-3	GN-4	GN-5	GN-6	GN-7	GN-8	GN-9	GN-10	GN-11	GN-12	GN-13	GN-14
Microbe	RWKRWWRLL	RWKRWWRWI	KWWRWRRWW	RWKKWWRWL	KKRWWWWWR	RKRWWWWFR	KWKRWWWFR	IWKRWWWKR	RIWKIWWKR	IKKRWWWFR	IKWKRWWWR	KWWKIWRWR	RIWKRWWFR	RLWKRWWIR
Pseudomonas aeruginosa (Gram-negative)														
H103 (wild type)	32	4	32	2	4	2	32	>128	64	128	128	32	>128	>128
213 (MDR)	64	32	128	16	32	8	>128	>128	>128	>128	>128	128	>128	>128
LES400 (MDR)	32	32	64	16	32	32	64	>128	128	>128	>128	64	>128	>128
Escherichia coli (Gram-negative)														
63103 (ESBL)	16	16	>128	16	8	8	16	64	64	64	32	64	64	64
Staphylococcus aureus (Gram-positive)														
ATCC25923	16	8	8	4	4	4	32	128	128	32	32	16	64	128
C623 (MRSA)	8	4	8	4	2	2	16	64	64	32	32	16	32	64
Enterococcus faecium (Gram-positive)														
mic80 (VRE; VanA)	8	32	32	16	8	8	32	128	32	64	32	32	64	64
Candida albicans (yeast)														
C627	8	4	8	4	4	8	32	32	8	128	8	8	128	16

P. aeruginosa PAO1 strain H103 (23), *S. aureus* ATCC#25923 (23). An methicillin-resistant *Staphylococcus aureus* (MRSA) clinical isolate (C623; Anthony Chow, Vancouver General Hospital, Vancouver, Canada). Vancomycin-resistant clinical isolates of *E. faecium* (Ana M. Paccagnella, BC Centre for Disease Control, Vancouver, Canada). Clinical isolate of *E. coli* expressing extended spectrum β-lactamases (ESBL, George Zhanel, Health Sciences Centre, Winnipeg, Canada). A clinical isolate (#213) of multidrug-resistant *P. aeruginosa* (Carlos Kiffer, University of São Paulo, Brazil). *P. aeruginosa* clinical isolates of the Liverpool epidemic strain (LES400) (24) (Craig Winstanley, University of Liverpool, UK). *C. albicans* (Barbara Dill, UBC Department Microbiology, Vancouver, Canada). See Materials and Methods for details.

B (codes GN-8 to -14), shown in Table 5. The MIC of wild-type *P. aeruginosa* strain H1001 used in the luminescence assay was 50–64 μg/mL for Bac2A (22). As *P. aeruginosa* strain H103 is the parent of H1001, we assume a similar MIC, and the value 32 μg/mL is close to the 50% of the Bac2A MIC (25 μg/mL). Therefore, MIC values for peptides GN-1 to GN-7 in simulation A were consistent with the criterion for high activity used previously. Only one of the seven peptides from simulation B had MIC $\leq$32 μg/mL. The activity of peptides varied depending on the microbe being treated but some peptides like GN-2, -4, -5, and -6 demonstrated excellent activity. These results suggest a bias attributed to starting peptide populations and the particular simulated evolution that led to the final populations.

Conclusions

We have described here the use of a genetic algorithm to efficiently identify novel peptides that have a high likelihood of being strongly antibacterial. In our previous studies, we created software models using ANNs that were found to be up to 94% accurate in predicting highly active peptides, when using a very large *in silico* library of 100 000 biased random sequences to identify additional peptides. In the current study, we demonstrated that the heuristic search method of GA identifies additional active peptides with considerably greater efficiency (0.50% of evaluated peptides) than our previous study with biased random sequences (0.026% of evaluated peptides). Currently, we evaluate QSAR descriptors for each peptide using commercial software on a small number of computers, a situation that strongly limits the number of peptides that can be evaluated. Hence, we find that the increased efficiency of the genetic algorithm methods allows a dramatically increased capability to identify novel antimicrobial peptide candidates. Nevertheless, we find that the activity of peptides that result from sequence searches using GA is strongly dependent on the initial starting population of peptides, despite the final fitness scores for the two simulations starting from similar distribution of amino acids being statistically similar. Based on leverage calculations for peptides, the most reliable predictions will be obtained for peptides most similar to those whose antibacterial activity was measured during the initial model construction. A more effective strategy would be to adopt an iterative approach, wherein computational and experimental approaches were used to identify new improved starting point for initiation of genetic algorithms, followed by retraining of the machine learning algorithms using the new data to improve the ability to predict peptide activity. Despite these restrictions on generality of prediction, we have reported here several novel peptides that are active against pathogens of clinical importance.

Acknowledgments

This manuscript is dedicated to the memory of Aaron W.J. Wyatt who tragically passed away on December 24, 2008. Aaron was not only a superb colleague but also a good friend. We gratefully acknowledge financial support from the Canadian Institutes for Health Research (CIHR) and the Advanced Foods and Materials Network and the Foundation of the National Institutes of Health and CIHR through the Grand Challenges in Global Health Initiative. RH is the recipient of a Canada Research Chair. CDF received a Doctoral Research Award from the CIHR. We also wish to acknowledge the helpful comments made by the reviewers that led to this improved manuscript.

References

1. Levy S.B., Marshall B. (2004) Antibacterial resistance worldwide: causes, challenges and responses. Nat Med;10:S122–S129.
2. Yeaman M.R., Yount N.Y. (2003) Mechanisms of antimicrobial peptide action and resistance. Pharmacol Rev;55:27–55.
3. Hamilton-Miller J.M.T. (2004) Antibiotic resistance from two perspectives: man and microbe. Int J Antimicrob Agents;23:209–212.
4. Koczulla A.R., Bals R. (2003) Antimicrobial peptides: current status and therapeutic potential. Drugs;63:389–407.
5. Finlay B.B., Hancock R.E.W. (2004) Can innate immunity be enhanced to treat microbial infections? Nat Rev Microbiol;2:497–504.
6. Hancock R.E.W., Sahl H.G. (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol;24:1551–1557.
7. Jenssen H., Hamill P., Hancock R.E.W. (2006) Peptide antimicrobial agents. Clin Microbiol Rev;19:491–511.
8. Cherkasov A., Hilpert K., Jenssen H., Fjell C.D., Waldbrook M., Mullaly S.C., Volkmer R., Hancock R.E. (2009) Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. ACS Chem Biol;4:65–74.
9. Fjell C.D., Jenssen H., Hilpert K., Cheung W., Panté N., Hancock R.E.W., Cherkasov A. (2009) Identification of novel antibacterial peptides by chemoinformatics and machine learning. J Med Chem;52:2006–2015.
10. Cherkasov A. (2005) 'Inductive' descriptors. 10 successful years in QSAR. Curr Comput Aided Drug Des;1:21–42.
11. Cherkasov A. (2005) Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks. Int J Mol Sci;6:63–86.
12. Karakoc E., Sahinalp S.C., Cherkasov A. (2006) Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. J Chem Inf Model;46:2167–2182.
13. Karakoc E., Cherkasov A., Sahinalp S.C. (2006) Distance based algorithms for small biomolecule classification and structural similarity search. Bioinformatics;15:243–251.
14. Parrill A.L. (1996) Evolutionary and genetic methods in drug design. Drug Discov Today;1:514–521.
15. Niculescu S.P. (2003) Artificial neural networks and genetic algorithms in QSAR. J Mol Struct;622:71–83.
16. Solmajer T., Zupan J. (2004) Optimization algorithms and natural computing in drug discovery. Drug Discov Today;1:247–252.
17. Weaver D.C. (2004) Applying data mining techniques to library design, lead generation and lead optimization. Curr Opin Chem Biol;8:264–270.

18. Patel S., Stott I.P., Bhakoo M., Elliott P. (1998) Patenting computer-designed peptides. J Comput Aided Mol Des;12: 543–556.
19. Patel S., Stott I., Bhakoo M., Elliott P.. (2002) Patenting evolved bacterial peptides. In: Bentley P.J., Corne D.W., editors. Creative Evolutionary Systems. Pages 525–545. San Francisco: Morgan Kaufmann Publishers Inc.
20. Faraway J.J. (2005) Linear Models with R. New York: Chapman and Hall/CRC.
21. Hemmateenejad B., Yazdani M. (2009) QSPR models for half-wave reduction potential of steroids: a comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. Anal Chim Acta;634:27–35.
22. Hilpert K., Volkmer-Engert R., Walter T., Hancock R.E.W. (2005) High-throughput generation of small antibacterial peptides with improved activity. Nat Biotechnol;23:1008–1012.
23. Wu M., Hancock R.E. (1999) Improved derivatives of bactenecin, a cyclic dodecameric antimicrobial cationic peptide. Antimicrob Agents Chemother;43:1274–1276.
24. Salunkhe P., Smart C.H., Morgan J.A., Panagea S., Walshaw M.J., Hart C.A., Geffers R., Tummler B., Winstanley C. (2005) A cystic fibrosis epidemic strain of *Pseudomonas aeruginosa* displays enhanced virulence and antimicrobial resistance. J Bacteriol;187:4908–4920.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Evolution of peptide scores.

Figure S2. Evolution of peptide scores at early generations.

Figure S3. Evolution of peptide amino acid composition.

Figure S4. Structure of an artificial neural network.

Figure S5. Half-normal plots of leverage.

Table S1. Inductive and conventional molecular descriptors for the quantitative structure-activity relationship modeling.

Table S2. Peptide purity data.

Table S3. Amino acid frequency of original population.

Table S4. Simulation from random initial peptides.

Table S5. Leverage values for experimental peptides.

Table S6. Leverage values for simulated peptides.

Appendix S1. Methods and materials.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.