Chem Biol Drug Des 2007; **70:** 134–142

Research Article

doi: 10.1111/j.1747-0285.2007.00543.x

Evaluating Different Descriptors for Model Design of Antimicrobial Peptides with Enhanced Activity Toward *P. aeruginosa*

Håvard Jenssen[1], Tore Lejon[2], Kai Hilpert[1], Christopher D. Fjell[3], Artem Cherkasov[3] and Robert E. W. Hancock[1,*]

[1]Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2]Department of Chemistry, University of Tromsø, N-9037 Tromsø, Norway
[3]Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, BC V5Z 3 J5, Canada
**Corresponding author: Robert E. W. Hancock, bob@cmdr.ubc.ca*

The number of isolated drug-resistant pathogenic microbes has increased drastically over the past decades, demonstrating an urgent need for new therapeutic interventions. Antimicrobial peptides have for a long time been looked upon as an interesting template for drug optimization. However, the process of optimizing peptide antimicrobial activity and specificity, using large peptide libraries is both tedious and expensive. Here, we describe the construction of a mathematical model for prediction, prior to synthesis, of peptide antibacterial activity toward *Pseudomonas aeruginosa*. By use of novel descriptors quantifying the contact energy between neighboring amino acids in addition to a set of inductive and conventional quantitative structure–activity relationship descriptors, we are able to model the peptides antibacterial activity. Cross-correlation and optimization of the implemented descriptor values have enabled us to build a model (Bac2a-#2) that was able to correctly predict the activity of 84% of the tested peptides, within a twofold deviation window of the corresponding IC_{50} values, measured earlier. The predictive power, is an average of 10 submodels, each predicting the activity of 20 randomly excluded peptides, with a predictive success of 16.7 ± 1.6 peptides. The model has also been proven significantly more accurate than a simpler model (Bac2a- #1), where the inductive and conventional quantitative structure–activity relationship descriptors were excluded.

Key words: antimicrobial peptides, partial least square projections to latent structures, prediction of activity, *Pseudomonas aeruginosa*, quantitative structure–activity relationships, screening libraries

Received 12 April 2007, revised 24 June 2007 and accepted for publication 26 June 2007

The increasing use of antimicrobials has resulted in an increasing problem with drug-resistant bacterial and fungal pathogens (1,2). This has created the need for the discovery and design of new antimicrobial drugs. However, major difficulties have been experienced in discovering new chemical structures with low host toxicity and broad spectrum activity. We have suggested that cationic peptides serve as a good template for the design of such a new generation of antimicrobials (3).

Several cationic antimicrobial peptides have been demonstrated to be quite effective in killing a wide selection of bacterial and fungal pathogens, including *Pseudomonas aeruginosa,* which is the third leading cause of hospital-associated infections and, as a result of chronic lung infections, the leading cause of morbidity and mortality in cystic fibrosis patients. Such peptides were initially demonstrated to target the bacterial cytoplasmic membrane but it is now recognized that many peptides translocate across the membrane and interact with cytoplasmic targets (3,4). A variety of modes of action have been ascribed to these peptides, but it has proven difficult to relate these modes of action to particular peptide sequences, as small changes can drastically affect structure (3,5). Thus, it has proven challenging to systematically design and optimize new peptides with improved antimicrobial activity. Traditional design and optimization studies of peptides are also known to be expensive and time-consuming. However, production costs and the time required for evaluation of activity can be drastically reduced by synthesis of large peptide libraries on cellulose membranes and high-throughput antibacterial testing (6) as demonstrated through design of a single-substitution peptide libraries based on Bac2a, a linear peptide derivative of the 12-amino acid bovine neutrophil peptide bactenecin (6,7). Another approach that has been used in streamlined peptide design has been to develop mathematical models to explain and predict the peptide activities. We have earlier demonstrated the use of principal component analysis (PCA) to explain the biologic activity of antimicrobial peptides (8,9). Partial least squares projection to latent structures (PLS) is another technique that has been used to build statistical models that can predict peptide activity prior to synthesis (10). A peptide library containing more than 217 000 theoretical peptide sequences was screened for theoretical antimicrobial activity by the use of such a PLS technique, and the results were verified by synthesizing a limited number of these peptides and confirming their antimicrobial activity. However, larger number of peptides have never been predicted and verified in these earlier studies. Another problem with the PCA/PLS approaches are that peptides with the same amino acid content, but different primary sequences, may appear as identical peptides in the model.

In this study, we adapted the PCA/PLS approach and conventional amino acid descriptors (11) to explain the activity of different peptides generated in a large Bac2a peptide library (6). To inculcate primary sequence information into the models, we introduced a new descriptor, quantifying contact energies between neighboring amino acids (12). We also introduced computer-simulated parameters describing biophysical properties of the entire peptide (13) and some well characterized quantitative structure–activity relationship (QSAR) descriptors (as implemented within the MOE programs: *Molecular Operational Environment* v. 2006.10, by Chemical Computation Group Inc., Montreal, Canada, 2006), in an attempt to more easily distinguish between biologically active and inactive peptides. All of these new descriptors were optimized and/or cross-correlated to avoid the introduction of noise and/or highly correlating data in the model. The predictive power of the model was then evaluated by random exclusion of peptides from the data set, and the creation of new models with the remaining peptides to examine if these models built on subsets of peptides could predict the activity of the 200 excluded peptides.

Materials and Methods

Biologic data

The data from a single-substitution Bac2a-library containing 228 peptides was used in this study (6). The antibacterial inhibitory concentration IC_{50} of these peptides was evaluated using a luciferase-based assay with *P. aeruginosa* PAO1 strain H1001, containing a constitutively expressed luciferase gene (*luxCDABE).*

Mathematical approach

It has been previously demonstrated that each of the 20 natural coded amino acids can be described by three specific descriptor values dealing with their hydrophilicity or hydrophobicity (z_1), size (z_2) and charge-related properties (z_3) (Table 1; 11). These descriptors encompass sufficient information to make it possible to explain and compare different peptide sequences and their antibacterial, antiviral, and anticancer activity (8–10,14–16). In an attempt to incorporate information about the peptide primary sequence into our predictions, we introduced, as a descriptor, the use of contact energy between neighboring amino acids (Table 2; 12). In addition, a set of 50 inductive molecular QSAR descriptors (13) and a set of 27 conventional QSAR descriptors were implemented to further discriminate between the different peptides with almost identical amino acid compositions. Thus, each peptide in the screening library can be described by three descriptor values for each amino acid (Table 1) a series of contact energy descriptors for each pair of amino acids, in a sliding window fashion (Table 2), and 78 biophysical inductive and conventional QSAR descriptors (Table 3).

Principal component analysis is a multivariate projection method designed to extract and display systematic variation in a data matrix *X*, by transforming a large number of potentially correlated variables into a smaller number of definitive and uncorrelated (independent) variables called principal components. The main variability in the data are accounted for in the first principal component, and with progressively declining variability accounted in the second component, and so forth.

Table 1: Descriptor scales z_1, z_2, and z_3 for amino acid (11)

Amino acid		Z_1	Z_2	Z_3
Ala	A	0.07	−1.73	0.09
Cys	C	0.71	−0.97	4.13
Asp	D	3.64	1.13	2.36
Glu	E	3.08	0.39	−0.07
Phe	F	−4.92	1.30	0.45
Gly	G	2.23	−5.36	0.30
His	H	2.41	1.74	1.11
Ile	I	−4.44	−1.68	−1.03
Lys	K	2.84	1.41	−3.14
Leu	L	−4.19	−1.03	−0.98
Met	M	−2.49	−0.27	−0.41
Asn	N	3.22	1.45	0.84
Pro	P	−1.22	0.88	2.23
Gln	Q	2.18	0.53	−1.14
Arg	R	2.88	2.52	−3.44
Ser	S	1.96	−1.63	0.57
Thr	T	0.92	−2.09	−1.40
Val	V	−2.69	−2.53	−1.29
Trp	W	−4.75	3.65	0.85
Tyr	Y	−1.39	2.32	0.01

With PLS, the primary matrix is divided into two matrices, one containing the traditional *z*-descriptors (*X*-matrix) and one containing the peptide biologic activities, the contact energy descriptors and the inductive and conventional QSAR descriptors (*Y*-matrix). Correlations between these two matrices are then calculated, using the SIMCA-P 10.0 software package.

The antibacterial activity of the Bac2a library (6) was carefully analyzed using PLS modeling. Optimization of the PLS model required the identification of potential cross-correlations in the inductive and conventional QSAR descriptor sets. Thus, Pearson product moment correlations were calculated with the exclusion criteria >0.95 or <−0.95. The contribution of contact energies, and inductive and conventional QSAR descriptors to the PLS model could vary substantially from component to component, and accordingly the contribution level was investigated for the sum of the first six components, setting an exclusion level at <0.5.

Different peptide activity prediction models were constructed with either the contact energy descriptors, or the inductive and conventional QSAR descriptors, or all three. The theoretical quality of the optimized model was then evaluated. To confirm the predictive ability after descriptor optimization, 20 peptides were randomly excluded from the Bac2a library a total of 10 times, resulting in 10 new models, with the already optimized descriptor settings. These 10 new models where then used to predict the antibacterial activity of the excluded peptides.

Table 2: Contact energies given in dimensionless units (12)

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	−2.72	−3.57	−1.7	−1.51	−4.81	−2.31	−2.41	−4.58	−1.31	−4.91	−3.94	−1.84	−2.03	−1.89	−1.83	−2.01	−2.32	−4.04	−3.82	−3.36
Cys		−5.44	−2.41	−2.27	−5.80	−3.16	−3.6	−5.5	−1.95	−5.83	−4.99	−2.59	−3.07	−2.85	−2.57	−2.86	−3.11	−4.96	−4.95	−4.16
Asp			−1.21	−1.02	−3.48	−1.59	−2.32	−3.17	−1.68	−3.4	−2.57	−1.68	−1.33	−1.46	−2.29	−1.63	−1.8	−2.48	−2.84	−2.76
Glu				−0.91	−3.56	−1.22	−2.15	−3.27	−1.8	−3.59	−2.89	−1.51	−1.26	−1.42	−2.27	−1.48	−1.74	−2.67	−2.99	−2.79
Phe					−7.26	−4.13	−4.77	−6.84	−3.36	−7.28	−6.56	−3.75	−4.25	−4.1	−3.98	−4.02	−4.28	−6.29	−6.16	−5.66
Gly						−2.24	−2.15	−3.78	−1.15	−4.16	−3.39	−1.74	−1.87	−1.66	−1.72	−1.82	−2.08	−3.38	−3.42	−3.01
His							−3.05	−4.14	−1.35	−4.54	−3.98	−2.08	−2.25	−1.98	−2.16	−2.11	−2.42	−3.58	−3.98	−3.52
Ile								−6.54	−3.01	−7.04	−6.02	−3.24	−3.76	−3.67	−3.63	−3.52	−4.03	−6.05	−5.78	−5.25
Lys									−0.12	−3.37	−2.48	−1.21	−0.97	−1.29	−0.59	−1.05	−1.31	−2.49	−2.69	−2.6
Leu										−7.37	−6.41	−3.74	−4.2	−4.04	−4.03	−3.92	−4.34	−6.48	−6.14	−5.67
Met											−5.46	−2.95	−3.45	−3.3	−3.12	−3.03	−3.51	−5.32	−5.55	−4.91
Asn												−1.68	−1.53	−1.71	−1.64	−1.58	−1.88	−2.83	−3.07	−2.76
Pro													−1.75	−1.73	−1.7	−1.57	−1.9	−3.32	−3.73	−3.19
Gln														−1.54	−1.8	−1.49	−1.9	−3.07	−3.11	−2.97
Arg															−1.55	−1.62	−1.9	−3.07	−3.41	−3.16
Ser																−1.67	−1.96	−3.05	−2.99	−2.78
Thr																	−2.12	−3.46	−3.22	−3.01
Val																		−5.52	−5.18	−4.62
Trp																			−5.06	−4.66
Tyr																				−4.17

Loading plots from the models were examined and used to evaluate the importance of the different amino acids, and identification of the theoretically most optimal amino acid compositions. The sequences of this theoretical peptide were then compared with multiple-substitution Bac2a analogs with known antibacterial activity (6).

Software

The program package SIMCA-P 10.0 from Umetrics (Umeå, Sweden) was used for PCA/PLS calculations. The theoretically derived amino acid descriptors (z_1–z_3) were centered prior to calculations, while the biologic activity and the remaining descriptors were all scaled to unit variance, to ensure they had equal influence in the model. General two-tailed *t*-test and non-parametric test (paired *t*-test) used to confirm statistically significant difference between the predictive power of Bac2a- #1 and #2 was calculated using PRISM® (GraphPad Software Inc., version 3.0, San Diego, CA, USA).

Results and Discussion

Over the past years, hundreds of peptides with sequences related to the naturally occurring host defense peptide bactenecin have been made, in an attempt to understand sequence requirements for antibacterial activity. The basis for the current studies was a single-substitution library based on Bac2a (6) containing 228 different peptides (Table 4).

The peptide library was synthesized on a cellulose membrane (17,18) and the peptide antibacterial activities against *P. aeruginosa* were analyzed with a luciferase-based killing assay (6). This reduced the cost of peptide synthesis in addition to providing a high-throughput screening assay for peptide antibacterial activity. However, to optimize this even further we attempted to create a computer simulation model able to predict peptide biologic activities prior to synthesis. Earlier work has taught us that specific amino acid descriptors (11) can be used to explain and predict a variety of different biologic peptide activities (8–10,14–16). However, such an approach with these descriptors was clearly not informative for the Bac2a library, resulting in only a single significant component (Table 5). This can be explained based on the primary sequence of the starting peptide, Bac2a (R_1LARIVVIRVAR_{12}). Bac2a is composed of only five different amino acids, and thus single substitutions in quite a few positions will create peptides with different primary sequences, but the same amino acid content, e.g. the same amino acid single substitution of R_1, R_4, R_9, or R_{12} will result in four peptides with different primary structure, but the same overall amino acid content. By performing PLS modeling with only the specific amino acid descriptors, these peptides with identical amino acid content would be interpreted as identical, even though their biologic activities in many cases are different, thus making it impossible to build a good model.

Contact energy descriptor model

By introducing amino acid contact energy descriptors, one value for each pair of amino acids, in a sliding window fashion, it was possible to distinguish between the peptides with identical amino acid contents but different sequences. It should be mentioned that these contact energy descriptors are calculated from a large selection of proteins and then transformed to reflect averaged interactions between specific types of amino acids. Thus, they may give a skewed reflection of the amino acid contact energies occurring in peptides, and their meaningful contribution in PLS modeling of peptides may vary. However, by using these values to describe the peptides in the Bac2a library, we are adding primary structure information to the PLS model, resulting in a model with 16 significant components, explaining 78% and 84% of the variation in the *X*- and *Y*-matrices, respectively (cross-validation Q2 = 79%; Table 5). In the process of optimization of the contact energy descriptors, examining their relative contribution to the model in the first six

Table 3: Inductive and conventional QSAR descriptors

Descriptor	Characterization	I/C	I/E
a acc	Number of hydrogen bond acceptor atoms	C	E
a don	Number of hydrogen bond donor atoms	C	E
ASA	Water accessible surface area	C	I
ASA−	Water accessible surface area of all atoms with negative partial charge	C	E
ASA H	Water accessible surface area of all hydrophobic atoms	C	I
ASA P	Water accessible surface area of all polar atoms	C	I
ASA+	Water accessible surface area of all atoms with positive partial charge	C	E
Average EO neg	Arithmetic mean of electronegativities of atoms with negative partial charge	I	E
Average EO pos	Arithmetic mean of electronegativities of atoms with positive partial charge	I	E
Average hardness	Arithmetic mean of hardnesses of all atoms of a molecule	I	E
Average neg charge	Arithmetic mean of negative partial charges on atoms of a molecule	I	E
Average neg hardness	Arithmetic mean of hardnesses of atoms with negative partial charge	I	I
Average neg softness	Arithmetic mean of softnesses of atoms with negative partial charge	I	E
Average pos charge	Arithmetic mean of positive partial charges on atoms of a molecule	I	I
Average pos hardness	Arithmetic mean of hardnesses of atoms with positive partial charge	I	I
Average pos softness	Arithmetic mean of softnesses of atoms with positive partial charge	I	E
Average softness	Arithmetic mean of softnesses of all atoms of a molecule	I	E
b 1rotN	Number of rotatable single bonds	C	I
EO equalized	Iteratively equalized electronegativity of a molecule	I	E
FCharge	Total charge of the molecule	C	I
Global hardness	Molecular hardness – reversed softness of a molecule	I	I
Global softness	Molecular softness – sum of constituent atomic softnesses	I	I
Hardness of most neg	Atomic hardness of an atom with the most negative charge	I	E
Hardness of most pos	Atomic hardness of an atom with the most positive charge	I	E
Largest neg hardness	Largest atomic hardness among values for negatively charged atoms	I	I
Largest neg softness	Largest atomic softness among values for positively charged atoms	I	E
Largest pos hardness	Largest atomic hardness among values for positively charged atoms	I	I
Largest pos softness	Largest atomic softness among values for positively charged atoms	I	E
Largest Rs i mol	Largest value of atomic steric influence *Rs* (*atom* → *molecule*) in a molecule	I	E
Largest Rs mol i	Largest value of steric influence *Rs* (*molecule* → *atom*) in a molecule	I	E
logP (o/w)	Log of the octanol/water partition coefficient	C	I
logS	Log of the aqueous solubility	C	I
Most neg charge	Largest partial charge among values for negatively charged atoms	I	E
Most neg Rs i mol	Steric influence *Rs* (*atom* → *molecule*) OF the most negatively charged atom to the rest of a molecule	I	E
Most neg Rs mol i	Steric influence *Rs* (*molecule* → *atom*) ON the most negatively charged atom in a molecule	I	E
Most neg Sigma i mol	Largest negative atomic inductive parameter σ^* (*atom* → *molecule*) for atoms in a molecule	I	E
Most neg Sigma mol i	Largest (by absolute value) negative group inductive parameter σ^* (*molecule* → *atom*) for atoms in a molecule	I	E
Most pos charge	Largest partial charge among values for positively charged atoms	I	E
Most pos Rs i mol	Steric influence *Rs* (*atom* → *molecule*) OF the most positively charged atom to the rest of a molecule	I	E
Most pos Rs mol i	Steric influence *Rs* (*molecule* → *atom*) ON the most positively charged atom in a molecule	I	E
Most pos sigma i mol	Largest positive atomic inductive parameter σ^* (*atom* → *molecule*) for atoms in a molecule	I	E
Most pos sigma mol i	Largest positive group inductive parameter σ^* (*molecule* → *atom*) for atoms in a molecule	I	E
mr	Molecular refractivity	C	E
PC−	Total negative partial charge	C	I
PC+	Total positive partial charge	C	I
RPC−	Relative negative partial charge	C	E
RPC+	Relative positive partial charge	C	E
Smallest neg hardness	Smallest atomic hardness among values for negatively charged atoms	I	E
Smallest neg softness	Smallest atomic softness among values for negatively charged atoms	I	E
Smallest pos hardness	Smallest atomic hardness among values for positively charged atoms	I	E
Smallest pos softness	Smallest atomic softness among values for positively charged atoms	I	E

Table 3: (Continued)

Descriptor	Characterization	I/C	I/E
Smallest Rs i mol	Smallest value of atomic steric influence *Rs* (*atom* → *molecule*) in a molecule	I	E
Smallest Rs mol i	Smallest value of group steric influence *Rs* (*molecule* → *atom*) in a molecule	I	E
Softness of most neg	Atomic softness of an atom with the most negative charge	I	E
Softness of most pos	Atomic softness of an atom with the most positive charge	I	E
Sum hardness	Sum of hardnesses of atoms of a molecule	I	I
Sum neg hardness	Sum of hardnesses of atoms with negative partial charge	I	I
Sum neg sigma mol i	Sum of all negative group inductive parameters σ^* (*molecule* → *atom*) within a molecule	I	E
Sum pos hardness	Sum of hardnesses of atoms with positive partial charge	I	E
Sum pos Sigma mol i	Sum of all positive group inductive parameters σ^* (*molecule* → *atom*) within a molecule	I	I
Total Abs Sigma mol i	Sum of absolute values of group inductive parameters σ^* (*molecule* → *atom*) for all atoms within a molecule	I	E
Total charge	Sum of absolute values of partial charges on all atoms of a molecule	I	E
Total charge formal	Sum of charges on all atoms of a molecule (formal charge of a molecule)	I	E
Total neg softness	Sum of softnesses of atoms with negative partial charge	I	I
Total pos softness	Sum of softnesses of atoms with positive partial charge	I	E
Total sigma mol i	Sum of inductive parameters σ^* (*molecule* → *atom*) for all atoms within a molecule	I	I
TPSA	Polar surface area	C	E
vdw area	van der Waals surface area calculated using a connection table approximation	C	E
vdw vol	van der Waals volume calculated using a connection table approximation	C	E
vol	van der Waals volume calculated using a grid approximation	C	E
VSA	van der Waals surface area using polyhedral representation	C	E
vsa acc	Approximation to the sum of VDW surface areas of pure hydrogen bond acceptors	C	E
vsa acid	Approximation to the sum of VDW surface areas of acidic atoms	C	E
vsa base	Approximation to the sum of VDW surface areas of basic atoms	C	E
vsa don	Approximation to the sum of VDW surface areas of pure hydrogen bond donors	C	E
vsa hyd	Approximation to the sum of VDW surface areas of hydrophobic atoms	C	I
Weight	Molecular weight	C	I

C/I, (C) conventional or (I) inductive quantitative structure–activity relationship (QSAR) descriptors; I/E, (I) included or (E) excluded in the final Bac2a- # 2 model. The conventional QSAR descriptors are all obtained from Molecular Operational Environment, 2004, by Chemical Computation Group Inc., Montreal, while the inductive descriptors have been described by Cherkasov (13).

components, it was demonstrated that the parameter describing the contact between amino acids 3 and 4, and amino acids 11 and 12, did not contribute any useful information. Thus, the contact energy descriptors for these two positions were removed and an optimized model was built, increasing the cross-validation to 82% (Table 5). The lack of information contributed by the contact energy descriptors in these two positions could be explained by the identical alanine–arginine patterns. Naturally, descriptors dealing with contact energy fall short when identical sequence patterns appear in any given peptide sequence, and approaches are currently being developed to resolve this issue.

Inductive and conventional QSAR descriptor model

The inductive and conventional QSAR descriptors were similarly used to distinguish between the peptides with identical amino acid content, and these descriptors were used in a new model (Table 4). Both the inductive and the conventional QSAR descriptors are molecular biophysical descriptors, thus the chance of cross-correlation between these two descriptor sets was considered rather high. To avoid the implementation of strongly correlating data into the model, with the potential for causing bias, these descriptor sets were evaluated using Pearson's correlation. This reduced the total number of descriptors from 87 in total, to 26 inductive and 17 conventional QSAR descriptors, resulting in a model that utilized 61% of *X*-matrix to model 58% and predict 4% of *Y*-matrix (Table 5). When investigating the contribution from the different inductive and conventional QSAR descriptors it became clear that several of them did not contribute significant information to the first six principal components. Removal of these less important descriptors (during optimization) resulted in a drastic increase in the predictive potential of the model, explaining 38% and 74% of the variation in the *X*- and *Y*-matrix, respectively (cross-validation Q2 = 55%; Table 5).

Combined model using contact energy, inductive and conventional QSAR descriptors

In the final modeling attempt we used the cross-correlated and optimized descriptor sets obtained as described above. This data set resulted in 17 significant components, explaining 78% and 82% of the variation in the *X*- and *Y*-matrix, respectively (cross-validation Q2 = 65%; Table 5). If only the final R2*Y* and Q2cum values are taken into account, this model appears weaker than the model

Table 4: The first two columns give the position (row number) and the one-letter code sequence of the native peptide. Second and third rows indicate column number and the amino acid substituted at each amino acid position

		Substituted amino acid																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Original amino acid		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	R	0.15	0.27	0.34	0.41	0.18	0.15	0.24	0.17	0.14	0.16	0.35	0.34	0.17	0.33	0.13	0.29	0.25	0.24	0.06	0.20
2	L	0.17	0.08	0.33	0.21	0.13	0.06	0.10	0.12	0.06	0.13	0.18	0.18	0.15	0.16	0.05	0.10	0.17	0.12	0.06	0.09
3	A	0.13	0.09	0.18	0.16	0.04	0.12	0.09	0.07	0.05	0.09	0.14	0.12	0.14	0.09	0.03	0.17	0.11	0.15	0.04	0.11
4	R	0.31	0.35	0.49	0.41	0.45	0.46	0.35	0.59	0.11	0.75	0.39	0.28	0.25	0.26	0.13	0.30	0.31	0.27	0.25	0.29
5	I	0.31	0.05	0.42	0.30	0.29	0.26	0.23	0.13	0.10	0.21	0.23	0.26	0.33	0.26	0.08	0.22	0.20	0.17	0.06	0.20
6	V	0.25	0.06	0.44	0.47	0.09	0.43	0.20	0.20	0.12	0.20	0.26	0.30	0.75	0.25	0.15	0.29	0.23	0.13	0.07	0.13
7	V	0.37	0.06	0.25	0.20	0.17	0.17	0.09	0.05	0.03	0.11	0.20	0.10	0.60	0.05	0.05	0.26	0.09	0.13	0.19	0.11
8	I	0.48	0.06	0.75	0.75	0.14	0.50	0.23	0.13	0.15	0.18	0.38	0.40	0.31	0.31	0.16	0.42	0.52	0.13	0.16	0.16
9	R	0.39	0.09	0.75	0.75	0.38	0.23	0.41	0.48	0.18	0.41	0.27	No fit	0.41	0.40	0.13	0.41	0.31	0.49	0.22	0.19
10	V	0.61	0.06	0.75	0.75	0.21	0.39	0.21	0.11	0.14	0.16	0.22	0.29	0.22	0.41	0.18	0.41	0.33	0.13	0.08	0.13
11	A	0.13	0.04	0.21	0.23	0.12	0.08	0.08	0.06	0.06	0.08	0.09	0.12	0.18	0.13	0.05	0.10	0.10	0.10	0.13	0.07
12	R	0.38	0.75	0.75	0.75	0.75	0.33	0.20	0.27	0.19	0.47	0.47	0.25	0.42	0.32	0.13	0.40	0.29	0.37	0.75	0.54

The matrix in the lower right corner (20 × 12) represents the 228 individual substitution peptides, and the values assigned to each peptide represent the peptides IC_{50} activity. The peptides are color coded in respect to the antibacterial activity that they possess; black indicating active peptides (IC_{50} ≤0.1), light grey are intermediate activity peptides (IC_{50} between 0.1 and 0.59) and white are classified as inactive peptides (IC_{50} ≥0.6) (6).

Table 5: The *X*-matrix in all the models contains the *z*-scale descriptors, while the content of the *Y*-matrix is pending on the different models, implementing contact energy (CE) and/or inductive and conventional QSAR descriptors (QSAR) both prior to and after optimization (opt.)

Model	*Y*-matrix	Comp.	R2*X*	R2*Y*	Q2cum
Bac2a- #1	IC_{50}	1	7.1	37.6	27.9
	CE + IC_{50}	16	78.4	83.7	78.9
	CE (opt.) + IC_{50}	14	74.1	86.5	81.7
Bac2a- #2	QSAR + IC_{50}	12	61.1	61.3	18.3
	QSAR (cross-correlated) + IC_{50}	12	61.0	57.9	3.8
	QSAR (cross-correlated & opt.) + IC_{50}	7	38.1	74.0	55.1
	CE (opt.) + QSAR (cross-correlated & opt.) + IC_{50}	17	78.0	81.5	65.4

Comp. is the number of significant components; R2*X* and R2*Y*: the fraction of the sum of squares of all the *X*'s and *Y*'s explained by the current component, respectively; Q2cum: the cumulative Q2 for the extracted components; QSAR, quantitative structure–activity relationship.

where only the contact energy descriptors were used. However, in the model using only the contact energy descriptors (Bac2a- #1), it required six components to reach a R2*Y* level of 50%, while this level was reached with only two components in the model using the combined descriptor set (Bac2a- #2). The reason for one model needing fewer components than another is that the data becomes 'stretched out' in one (or a few) dimension(s). Thus, the descriptors used in Bac2a- #2 better described the variation between the objects (peptides) than did the descriptors used in Bac2a- #1, indicating that Bac2a- #2 may be the best model. Given these rather conflicting results, both models were evaluated further, to assure that the most accurate one was chosen.

Prediction of antibacterial peptides

To confirm the predictive capacity of the Bac2a- #1 and #2, we randomly extracted 20 peptides from the Bac2a-library 10 times, and built models on the remaining 208 peptides using both the Bac2a- #1 and #2 optimized descriptor settings, resulting in 20 new models. The extraction of peptides was truly random, resulting in minimal overlap between the peptides extracted from each model (7.5 ± 2.5%). The Bac2a- #1-related models resulted in an average of 14.1 significant components, utilizing 74.4% of *X* (R2*X* = 0.744), modeling 86.1% of *Y* (R2*Y* = 0.861), with a cross-validation of 81.2% (Q2 = 0.812). Conversely, all of the Bac2a- #2-related models had 17 significant components, explaining an average of 78.2% of *X* (R2*X* = 0.782) and 81.2% of *Y* (R2*Y* = 0.812) with a cross-validation of 65.6% (Q2 = 0.656). These models where then used to predict the antibacterial activity of the randomly excluded peptides, and these results were compared to their measured antibacterial activities. Given the nature of the assay measuring the peptide IC_{50} values, we allowed a twofold deviation window in antibacterial activity when evaluating the success of the predictive model. The results demonstrated that the antibacterial activity were predicted correctly for an average of 15.1 ± 2.3 (76%) and 16.7 ± 1.6 (84%) of the 20 peptides excluded in each model related to the Bac2a- #1 and #2, respectively. A *t*-test (and non-parametric test) with paired testing of predictive results for all 10 submodels of Bac2a- #1 and #2 with a confidence interval of 99%, resulted in a two-tailed p-value of 0.0084, confirming that the two models are statistically significantly different. This suggests that the model built with all the optimized and cross-correlated descriptors (Bac2a- #2) gives a better and more accurate representation of the peptide structure/activity relationships, than the model only implementing the contact energy descriptors (Bac2a- #1). In addition, a higher spread in the peptide-predicted activities within the twofold window was observed when using the Bac2a- #1- compared to the Bac2a- #2-related models (data not shown). Evaluation of incorrectly predicted peptides by use of Bac2a- #1 and Bac2a- #2 with a paired *t*-test, demonstrated that both models have a statistically even spread of peptides predicted with higher and lower activity compared to the observed IC_{50} values, although the numerical data may indicate that a higher percentage of the peptides are predicted with lower activity than observed (data not shown).

It can be anticipated, because of the nature of PLS modeling, that peptides at either ends of the antibacterial activity range would be predicted less accurately than the ones in the middle of the scale. However, when examining the group of peptides that were not successfully predicted, we found nine peptides with superior activity, eight intermediately active and four inactive peptides, having IC_{50} values of <0.1, between 0.1 and 0.59, and ≥0.6, respectively. This may indicate that the model could have a higher chance of predicting false-positive (active peptides) than predicting false-negative (inactive peptides).

The loading plot of the Bac2a- #2 model, demonstrated the relative importance of the different variables in the model (Figure 1). Thus, variables originated far from the origin were most important for the modeling. The loading plot also revealed that there is a co-variance between the peptide antibacterial activities and several different biophysical parameters. Systematic evaluation of the loading plot gave a ranking of preferred amino acids located in the different positions throughout the peptide sequence. This type of evaluation has been demonstrated quite successful in explaining why some peptides are more active than others (8,9).

Peptide optimization

A brief evaluation of the loading plot confirms that the most important amino acids are in positions 1, 4, 9, and 12, and they should all preferably be arginine residues (Table 6). In positions 3 and 11, the residue introduced should optimally be a tryptophan, arginine, or tyrosine, with the latter being the least favorable amino acid. Another improvement would be provided by the introduction of the same set of amino acids (W, R, or Y) at positions 6, 7, and 10, while modifications at positions 2, 5, and 8 would be less favorable. According to this rationale, we can explain the increase in activity from Bac2a to the multiple-substitution peptide Sub2 (Table 6; 6), as well as the increase in activity from the peptide Sub2, to Sub3 and Sub4, indicating that contact energy descriptors may enable prediction of the activity of structurally different peptides. These results also indicate that the arginine substitution at position 2 in these peptides might well limit their activity. Even further limitations are likely to be introduced by changing this arginine to a tryptophan in the peptide Sub6 (Table 6).

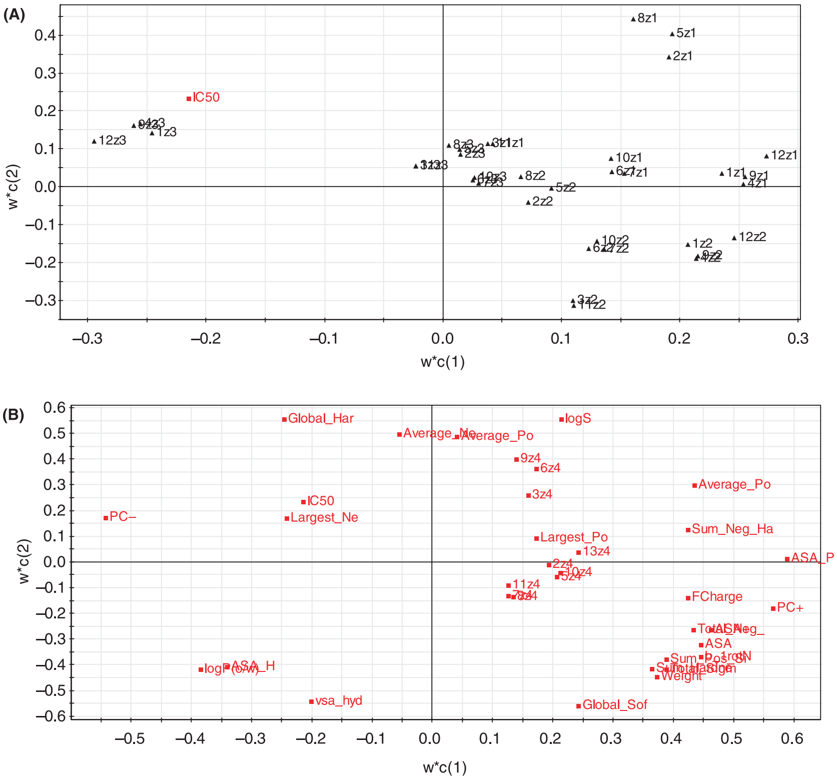

Figure 1: Loading plot for the Bac2a- #2 model using: (A) the *x*-variables including IC_{50}, (B) using only the *y*-variables.

Table 6: The small horizontal lines indicate that the amino acid in this position is identical as in Bac2a

	Primary peptide sequence												MIC (µg/mL)
Peptide	1	2	3	4	5	6	7	8	9	10	11	12	*Pseudomonas aeruginosa*
Bac2a	R	L	A	R	I	V	V	I	R	V	A	R	50
Loading plot predicted	–	–	W R Y	–	–	W R Y	W R Y	–	–	W R Y	W R Y	–	
Sub2	–	–	R	–	–	–	–	–	–	–	R	–	8
Sub3	–	R	W	–	–	–	–	–	–	–	R	–	2
Sub5	–	R	W	K	–	–	–	–	–	W	R	–	2
Sub6	–	W	W	K	–	W	–	–	–	W	W	–	31

The loading plot-predicted peptide is a theoretical peptide, and for some positions several amino acids are suggested with the most favorable on at the top. The antibacterial activity of Sub2, 3, 5, and 6 are published earlier (6).

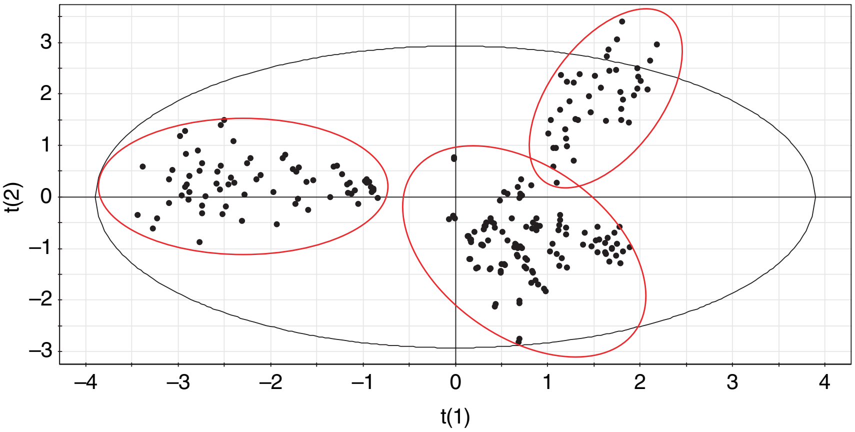

Figure 2: Score plot demonstrating subgrouping of the peptides into three groups: group 1 being to the left, group 2 in the middle, and group 3 in the upper right corner.

The optimal peptide sequence that is generated from the loading plot (Figure 1, Table 6; $R_1LXR_4I_5XXI_8R_9XXR_{12}$, where X represents the spectrum of alternative amino acids), can also be compared with the antibacterial activity results from the entire Bac2a library (Table 4). When simplifying the activity results from the library and only looking at different activity groups (black, grey, and white color code), it was evident that, with minor exceptions, changes at amino acid position 1, 4, 5, 8, 9, and 12 resulted in no significant positive changes in peptide antibacterial activity, thus confirming the results observed in the loading plot. Surprisingly, substitutions at position 2, appeared quite often to have a positive effect on antibacterial activity, and this was not picked up in the loading plot. Similarly, changes in positions 6 and 10, primarily led to minor changes in antibacterial activity, while the loading plot identified these as important positions to do substitutions. Changes in the remaining positions (3, 7, and 11) all gave several peptide candidates with increased antibacterial activity, consistent with results from the loading plot.

An additional finding during the modeling of Bac2a- #2 was the very distinct subgrouping of peptides in the score plot (Figure 2). In general, group 1 contained peptides with substitutions in sequence positions 1, 4, 9, and 12, group 2 had primarily substitutions in positions 3, 6, 7, 10, and 11, while group 3 had substitutions in positions 2, 5, and 8. Though it is beyond the scope of this paper to examine these subgroups in detail, this indicates that it might be possible to separate the Bac2a library into subgroups, which could be modeled better and more accurately than the entire library. This could be interesting to specific studies dealing with only optimization of certain parts of the peptide sequence.

Conclusions

We have demonstrated that contact energy descriptors can be used to implement information regarding the peptide primary sequence into PLS models, thus making it possible to distinguish closely related peptides with rather different antibacterial activities. We have also demonstrated that this model can be used to specifically predict the antibacterial activity of large sets of peptides, unknown prior to modeling. Comprising of Bac2a- #1 and - #2, demonstrated that both contact energy descriptors and inductive and conventional QSAR descriptors should be implemented in the model, to insure as precise prediction as possible. Similar modeling can be very useful in future high-throughput peptide design, reducing the number of peptides synthesized in search for the optimally active candidates.

Acknowledgments

We thank Timothy Chan and Avinask Chikatamarla for rewarding discussion. We gratefully acknowledge financial support through grants from the Canadian Institutes for Health Research (CIHR) and the Applied Food and Materials Network. KH received a CIHR postdoctoral fellowship. REWH was the recipient of a Canada Research Chair.

References

1. Levy S.B., Marshall B. (2004) Antibacterial resistance worldwide: causes, challenges and responses. Nat Med;10(Suppl. 12):S122–S129.
2. Overbye K.M., Barrett J.F. (2005) Antibiotics: where did we go wrong? Drug Discov Today;10:45–52.
3. Hancock R.E.W., Sahl H.G. (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol;24:1551–1557.
4. Hancock R.E.W., Rozek A. (2002) Role of membranes in the activities of antimicrobial cationic peptides. FEMS Microbiol Lett;206:143–149.
5. Jenssen H., Hamill P., Hancock R.E.W. (2006) Peptide antimicrobial agents. Clin Microbiol Rev;19:491–511.
6. 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.
7. Romeo D., Skerlavaj B., Bolognesi M., Gennaro R. (1988) Structure and bactericidal activity of an antibiotic dodecapeptide purified from bovine neutrophils. J Biol Chem;263:9573–9575.
8. Jenssen H., Gutteberg T.J., Lejon T. (2005) Modelling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. J Pept Sci;11:97–103.
9. Jenssen H., Gutteberg T.J., Lejon T. (2005) Modelling the anti-herpes simplex virus activity of small cationic peptides using amino acid descriptors. J Pept Res;66(Suppl. 1):48–56.
10. Jenssen H., Gutteberg T.J., Rekdal O., Lejon T. (2006) Prediction of activity, synthesis and biological testing of anti-HSV active peptides. Chem Biol Drug Des;68:58–66.
11. Hellberg S., Sjostrom M., Skagerberg B., Wold S. (1987) Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem;30:1126–1135.
12. Miyazawa S., Jernigan R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol;256:623–644.
13. Cherkasov A. (2005) Inductive QSAR descriptors, distinguishing compounds with antibacterial activity by artificial neural network. Int J Mol Sci;6:63–86.
14. Lejon T., Stiberg T., Strom M.B., Svendsen J.S. (2004) Prediction of antibiotic activity and synthesis of new pentadecapeptides based on lactoferricins. J Pept Sci;10:329–335.
15. Lejon T., Strom M.B., Svendsen J.S. (2001) Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. J Pept Sci;7:74–81.
16. Yang N., Lejon T., Rekdal O. (2003) Antitumour activity and specificity as a function of substitutions in the lipophilic sector of helical lactoferrin-derived peptide. J Pept Sci;9:300–311.
17. Frank R. (1992) Spot synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. Tetrahedron;48:9217–9232.
18. Frank R. (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports – principles and applications. J Immunol Methods;267:13–26.