Journal of Peptide Science
J. Pept. Sci. 2008; **14**: 110–114
Published online 10 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/psc.908

JOURNAL OF PEPTIDE SCIENCE

Short Communication

QSAR modeling and computer-aided design of antimicrobial peptides

HÅVARD JENSSEN,[a] CHRISTOPHER D. FJELL,[b] ARTEM CHERKASOV[b] and ROBERT E. W. HANCOCK[a]*

[a] Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
[b] Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, BC, V5Z 3J5, Canada

Received 13 June 2007; Accepted 24 June 2007

Abstract: The drastic increase in multi-drug-resistant bacteria has created an urgent need for new therapeutic interventions, including antimicrobial peptides, an interesting template for novel drug development. However, the process of optimizing peptide antimicrobial activity and specificity using large peptide libraries is both tedious and expensive. Here we confirm the use of a mathematical model for prediction, prior to synthesis, of peptide antibacterial activity toward the antibiotic resistant pathogen *Pseudomonas aeruginosa*. By the use of novel descriptors quantifying the contact energy between neighboring amino acids, as well as a set of inductive and conventional QSAR descriptors, we were able to model the antibacterial activity of peptides. Cross-correlation and optimization of the implemented descriptor values enabled us to build two models, using very limited sets of peptides, which were able to correctly predict the activity of 85 or 71% of the tested peptides, within a twofold deviation window of the corresponding previously assessed IC_{50} values, measured earlier. Though these two models were significantly different in size, they demonstrated no significant difference in their predictive power, implying that it is possible to build powerful predictive models using even small sets of structurally different peptides, when using contact-energy descriptors and inductive and conventional QSAR descriptors in the model design. Copyright © 2007 European Peptide Society and John Wiley & Sons, Ltd.

Keywords: *P. aeruginosa*; antimicrobial peptides; quantitative structure-activity relationships; prediction of activity; partial least square projections to latent structures; screening libraries

INTRODUCTION

Bacterial resistance has increased dramatically over the past decade [1], presenting a huge global health threat and a challenge for antimicrobial drug developers [2]. Although small molecules still dominate drug discovery, peptides have recently been recognized as suitable leads in several areas of drug discovery owing to their high affinity and specificity toward their targets. Importantly, the toxicity profiles of peptide-based therapeutics are usually very favorable [3]. At the same time, the potential rapid renal clearance and poor *in vivo* stability of peptides resulting from protease degradation can contribute to their short half-life. Taken together with the suggestion of low bioavailability, these arguments have been typically used to argue against peptide lead development. However, Scott *et al.* [4] recently reported an innate defense-regulator peptide that sustained immunomodulatory activity in an *in vivo* model for 54 h. In addition, peptide-based antimicrobials are without doubt suitable for topical applications [5], offering a decreased potential for resistance induction [6] compared to other antimicrobials.

Large-scale screening projects have involved the screening of both naturally occurring peptides and chemical and genetic/recombinant peptide libraries in the search for new lead molecules [7,8]. Unfortunately such manual or semiautomated techniques are quite labor intensive and expensive. Computer-aided predictions of peptide antimicrobial activity using soft independent modeling of class analogy (SIMCA), and incorporated principal component analysis (PCA)/partial least squares projection to latent structures (PLS) algorithms, have also demonstrated some success [9–13]. However, a persistent problem with this type of mathematical modeling has been that no primary structure information has been implemented in the models, thereby preventing the effective analysis of peptides with large structural diversities. We recently attempted to solve this problem by introducing contact energy between neighboring amino acids [14] as a descriptor. Although this ignores all intermolecular interactions involved in determining three-dimensional structure, except for those between neighboring amino acids in the primary structure, this implementation resulted in a rather powerful predictive model [15]. However, concerns regarding the use of such contact-energy descriptors remain in an over-simplified fashion. Therefore, in this paper we examine the value of using contact-energy

*Correspondence to: Robert E. W. Hancock, Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; e-mail: bob@cmdr.ubc.ca

descriptors in the design of PLS models for antimicrobial peptides. The robustness of these descriptors in combination with inductive and conventional quantitative struture–activity relationship (QSAR) descriptors [16] is demonstrated here by using only a small fraction of the peptide libraries for model building and then evaluating the predictive power of such models in estimating the activity of the remaining peptides. We also investigate whether these new descriptors are sufficient for building predictive models with one class of peptides to permit the prediction of the activity of structurally diverse antimicrobial peptides.

MATERIALS AND METHODS

Biological Data

The published data from two single substitution libraries (Bac034 and Bac2A), containing 216 and 228 peptides respectively, were used [17,18]. Peptide antibacterial activity was evaluated with a luciferase-based assay using *Pseudomonas aeruginosa* PAO1 strain H1001, measuring IC_{50} (or the related measurement EC_{50}, see below) concentrations, in addition to conventional minimal inhibitory concentration (MIC) analysis on selected peptides.

Mathematical Approach

The use of specific descriptor values dealing with amino acid hydrophilicity or hydrophobicity, size and charge-related properties [19] was as previously described, used for successful modeling of the antimicrobial activity of peptides [9–13,20]. Similarly, the implementation of primary structural information on peptides into QSAR models, by using contact energy between neighboring amino acids [14], as one of the descriptors was applied as described previously [15].

PCA and partial least squares PLS were done as described earlier [15] to analyze the structure–activity relationship of the peptides in the Bac034 library, and to verify the value of contact energy and inductive and conventional QSAR descriptors [16]. Merged and separate models of both the Bac034 and Bac2A libraries were made to investigate if they could be used to predict the biological activity of peptides with unrelated sequences and structures. The strength of these types of predictive models was evaluated by building stringent models on smaller fractions of the peptide library, predicting the activity of the peptides excluded from the model.

Software

The program package Simca-P 10.0 from Umetrics, Umeå, Sweden, was used for PCA/PLS calculations. The theoretically derived amino acid descriptors were centred prior to calculations, while the antibacterial activity and the remaining descriptors were all scaled to unit variance, to ensure that they had equal influence in the model. A Chi-squared test was used to confirm statistical differences between the predictive power of different sub-models of the 50-50 and 25-25 model, as calculated using PRISM (GraphPad Software Inc., version 3.0, San Diego, CA).

RESULTS AND DISCUSSION

Synthesis of large peptide libraries on cellulose membranes [21] has enabled us to investigate hundreds of peptides with sequences related to the naturally occurring host-defense peptide bactenecin, in an attempt to optimize and understand the determinants of antibacterial activity for these peptides and their mode of action. The basis for the current studies were two substitution libraries based on peptides Bac2A [18] and its scrambled sequence variant Bac034 [17], containing 216 and 228 different peptides, respectively.

Although cellulose libraries are relatively affordable, a considerable amount of work is warranted in understanding what makes a peptide active, since the ability to predict activity prior to synthesis would streamline peptide design. Part of this work can be done by statistical analysis and mathematical prediction modeling using PLS. Until recently PLS modeling of antimicrobial peptides solely utilized specific amino acid descriptors [19], limiting the modeling to peptides with significantly similar primary structures; but at the same time it required the overall amino acid composition of the peptides to be different. To illustrate this, we generated a model, containing all the peptides from the Bac034- and Bac2A-libraries, with only two significant components explaining 72 and 25% of the variation in the X- and Y-matrices, respectively (cross-validation $Q2 = 19\%$) (Table 1). However, we recently demonstrated that, by using amino acid contact-energy descriptors and a set of inductive and conventional QSAR descriptors, we were able to incorporate primary structure information into the modeling step for the Bac2A library, thereby significantly increasing the predictive ability of models and enabling the explanation of 78 and 82% of the variation in the X and Y matrices, respectively (cross-validation $Q2 = 65\%$) [15].

The amino acid contact-energy descriptors were derived from an average value of residue–residue contacts in a set of PDB-available proteins; therefore direct implementation of these values to the described peptides may not be entirely correct or precise. Questions remained regarding the accuracy in modeling the Bac2A library with contact-energy descriptors and whether there was a sequence-specific element to our earlier successes. Therefore we applied these strategies and descriptors to a different library (Bac034). By using the same cross-correlation and optimization steps in the model design as described earlier for the Bac2A library [15], we built a separate Bac034 model, explaining 86 and 80% of the variation in the X and Y matrices, respectively (cross-validation $Q2 = 57\%$) (Table 1). This confirms the value of using amino acid contact energy as a descriptor for this type of modeling experiment, and indicates that its success is not peptide-sequence dependent. To evaluate whether

DOI: 10.1002/psc

Table 1 The X-matrix in all the models contains the z-scale descriptors, while the content of the Y-matrix is dependent on the different models, implementing contact energy (CE) and/or inductive and conventional QSAR descriptors (QSAR), the latter being optimized and cross-correlated as described earlier [15]. Comp. is the number of significant components. R2X and R2Y are respectively the fractions of the sum of squares of all the X's and Y's explained by the current component, respectively. Q2 is the fraction of the total variation of the Y's that can be predicted by a component according to cross-validation, and Q2cum is the cumulative Q2 for the extracted components

Model	Y-matrix	Comp.	R2X	R2Y	Q2cum
Bac034 + Bac2A	$IC_{50} + EC_{50}$	2	72.1	24.7	18.9
Bac034	$CE + QSAR + EC_{50}$	21	86.1	79.9	57.3
Bac034 + Bac2A	$CE + QSAR + EC_{50} + IC_{50}$	17	92.4	82.2	76.9
50-50 model	$CE + QSAR + EC_{50} + IC_{50}$	17	92.4	82.5	77.3
50-50 model—Bac2A subset	$CE + QSAR + IC_{50}$	19	82.0	75.4	58.4
50-50 model—Bac034 subset	$CE + QSAR + EC_{50}$	16	78.4	77.8	60.0
25-25 model	$CE + QSAR + EC_{50} + IC_{50}$	17	92.3	84.2	79.1
25-25 model—Bac2A subset	$CE + QSAR + IC_{50}$	22	86.3	77.9	56.8
25-25 model—Bac034 subset	$CE + QSAR + EC_{50}$	20	84.7	80.5	58.8

the implementation of contact-energy descriptors or the inductive and conventional QSAR descriptors would enable modeling of structurally different peptides, an even more successful merged model of the two libraries was constructed, explaining 92 and 82% of the variation in the X- and Y-matrices, respectively (cross-validation Q2 = 77%) (Table 1). Though the score plot clearly divided the peptides from the two libraries into two subgroups (data not shown), it demonstrates the potential of building predictive models on peptides with diverse primary structures.

We hypothesized that by implementing primary structure descriptions of the peptides into the model it would be possible to build robust predictive models on a set of structurally diverse peptides, or use a model of the Bac034-library and predict the activity of the peptides in the Bac2A-library, or vice versa. To test this hypothesis it was crucial to have one common read-out from both libraries. IC_{50} is used in the lab as a screening tool for estimation of the MICs of peptides and there appears to be a fairly good correlation between these two parameters (unpublished results). EC_{50} was obtained for the Bac034 library in a similar way to IC_{50} for Bac2A, except that EC_{50} was calculated without experimentally derived, extrapolated baseline corrections; however, it was initially assumed that there should be a similar relationship between MIC and EC_{50}.

Thus a selected number of peptides tested for IC_{50}/EC_{50} and MIC were used to predict estimated MIC values for peptides in both libraries by the use of linear regression (data not shown). By attempting predictions separately for the Bac034 and Bac2A libraries, we were not able to extrapolate the derived relationship for one library to predict, with any significant accuracy, the activity of peptides in the other library. This probably reflects the above observation that these peptides clearly separated into two distinct classes, and may suggest that IC_{50} and EC_{50} are significantly different measurements (at least for the two libraries investigated). An approach to overcome this would be to include peptides from both libraries in a single, merged model. Therefore half the peptides from each library were randomly selected, combined and trained, using the optimal settings described earlier [15], in an attempt to probe the potential of such mixed peptide models (Table 1). The resultant 50-50 model had reasonable predictive power, demonstrating an 85.1% success in predicting the IC_{50} values of the excluded Bac2A peptides (Table 2), which is similar to the success rate of earlier Bac2A models [15].

In contrast, the 50-50 model had a relatively poor predictive success of 33.9% of the Bac034 library. The problem of predicting EC_{50} values may reflect on the rather inaccurate nature of the EC_{50} values, indicating that IC_{50} values, which incorporated background corrections estimated from the experimental datasets, are most likely a better way of assessing antibacterial activity.

The recently published model of the Bac2A library, developed by considering the activities of 90% of measured peptides, demonstrated a predictive power of 84% for the remaining randomly excluded 10% of the peptides. Though this randomization approach was repeated ten times, giving reproducible results, bias might have been introduced by using so many of the peptides in the data set to make predictions, and it could be argued that there is limited utility in predicting the activity of relatively few related peptides (i.e. 1 in 10). To address this, and to confirm that this mathematical approach would be useful with different datasets, a new model (25-25 model) was built using random selections of 25% of the peptides from both libraries for model-building and attempting to predict the remaining 75%. This model could explain 92%

DOI: 10.1002/psc

Table 2 Predictive power of the different models given in percentage correct predicted peptides, evaluated in respect to correctly predicted; IC_{50} and EC_{50} values among excluded peptides from the Bac2A and Bac034 library, respectively. Chi-square is the statistically significant difference (confidence intervals of 95%) between the main model (50-50 or 25-25) and the different submodels containing either only the Bac2A subset or the Bac034 subset

Prediction model	Bac2A		Bac034	
	IC_{50}	Chi-Square	EC_{50}	Chi-Square
50-50 model	85.1		33.9	
50-50 model—Bac2A subset	76.7	No $P = 0.2031$	—	Yes $P < 0.0001$
50-50 model—Bac034 subset	—		66.1	
25-25 model	70.7		28.2	
25-25 model—Bac2A subset	73.7	No $P = 0.7342$	—	Yes $P = 0.0008$
25-25 model—Bac034 subset	—		48.5	

of the variation in X- and 84% of the variation in the Y -matrix, (cross-validation = 79%) (Table 1). When examining its predictive ability, it became evident that 70.7% (121 peptides) of the excluded Bac2A peptides could be predicted correctly with respect to their measured IC_{50} values (Table 2). This drop in predictive power compared to the power of the earlier published Bac2A model [15] is probably a combined effect of substantially reducing the number of peptides utilized to develop the new model, and also the addition of a set of structurally different peptides (25% of the Bac034 library) to the model. Evaluation of the 25-25 model demonstrated the correct prediction of only 28.2% of the EC_{50} values (Table 2), confirming that IC_{50} values could be modeled with superior accuracy compared to EC_{50} values.

To further evaluate the effect of merging two peptide libraries in PLS modeling, new models were generated by excluding either the Bac034 or the Bac2A datasets from both the 50-50 model and the 25-25 model, resulting in Bac2A and Bac034 subsets of both the original models, respectively (Table 2). When comparing these new models with the original models, it is evident that Bac034 exclusion had no significant effect on the accuracy of the model. However, Bac2A exclusion resulted in a model with a significantly higher predictive power. This may be explained by IC_{50} being a highly accurate measurement, the inclusion of which will cause problems in predicting less accurate EC_{50} values, while exclusion of which will enable better modeling of the EC_{50} values. The results strengthen the conclusion that the success of a predictive model is highly influenced by the accuracy of the parameters included in the model design.

CONCLUSION

We have reconfirmed that contact-energy descriptors can be used for implementing primary structure information into PLS models, enabling modeling of structurally diverse peptides. We have also successfully built a predictive model on a limited selection of peptides from two distinctly related peptide libraries, and demonstrated its potential of correctly predicting the activity of the excluded part of the library.

Acknowledgements

We gratefully acknowledge financial support through grants from the Canadian Institutes for Health Research (CIHR) and the Applied Food and Materials Network. REWH was the recipient of a Canada Research Chair.

REFERENCES

1. Overbye KM, Barrett JF. Antibiotics: where did we go wrong? *Drug Discovery Today* 2005; **10**: 45–52.
2. Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. *Nat. Med.* 2004; **10**: S122–S129.
3. Dathe M, Schumann M, Wieprecht T, Winkler A, Beyermann M, Krause E, Matsuzaki K, Murase O, Bienert M. Peptide helicity and membrane surface charge modulate the balance of electrostatic and hydrophobic interactions with lipid bilayers and biological membranes. *Biochemistry* 1996; **35**: 12612–12622.
4. Scott MG, Dullaghan E, Mookherjee N, Glavas N, Waldbrook M, Thompson A, Wang A, Lee K, Doria S, Hamill P, Yu JJ, Li Y, Donini O, Guarna MM, Finlay BB, North JR, Hancock REW. An anti-infective peptide that selectively modulates the innate immune response. *Nat. Biotechnol.* 2007; **25**: 465–472.
5. Bush K, Macielag M, Weidner-Wells M. Taking inventory: antibacterial agents currently at or beyond phase 1. *Curr. Opin. Microbiol.* 2004; **7**: 466–476.

6. Ge Y, MacDonald DL, Holroyd KJ, Thornsberry C, Wexler H, Zasloff M. In vitro antibacterial properties of pexiganan, an analog of magainin. *Antimicrob. Agents Chemother.* 1999; **43**: 782–788.
7. Houghten RA, Pinilla C, Blondelle SE, Appel JR, Dooley CT, Cuervo JH. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* 1991; **354**: 84–86.
8. Watt PM. Screening for peptide drugs from the natural repertoire of biodiverse protein folds. *Nat. Biotechnol.* 2006; **24**: 177–183.
9. Jenssen H, Gutteberg TJ, Lejon T. Modelling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. *J. Pept. Sci.* 2005; **11**: 97–103.
10. Jenssen H, Gutteberg TJ, Rekdal O, Lejon T. Prediction of activity, synthesis and biological testing of anti-HSV active peptides. *Chem. Biol. Drug Des.* 2006; **68**: 58–66.
11. Lejon T, Stiberg T, Strom MB, Svendsen JS. Prediction of antibiotic activity and synthesis of new pentadecapeptides based on lactoferricins. *J. Pept. Sci.* 2004; **10**: 329–335.
12. Lejon T, Strom MB, Svendsen JS. Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J. Pept. Sci.* 2001; **7**: 74–81.
13. Yang N, Lejon T, Rekdal O. Antitumour activity and specificity as a function of substitutions in the lipophilic sector of helical lactoferrin-derived peptide. *J. Pept. Sci.* 2003; **9**: 300–311.
14. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 1996; **256**: 623–644.
15. Jenssen H, Lejon T, Hilpert K, Fjell C, Cherkasov A, Hancock REW. Evaluating different descriptors for model design of antimicrobial 12-mer peptides with enhanced activity towards P. aeruginosa. *Chem. Biol. Drug Des.* 2007; **70**: 134–142.
16. Cherkasov A. Inductive QSAR descriptors, distinguishing compounds with antibacterial activity by artificial neural network. *Int. J. Mol. Sci.* 2005; **6**: 63–86.
17. Hilpert K, Elliott MR, Volkmer-Engert R, Henklein P, Donini O, Zhou Q, Winkler DF, Hancock REW. Sequence requirements and an optimization strategy for short antimicrobial peptides. *Chem. Biol.* 2006; **13**: 1101–1107.
18. Hilpert K, Volkmer-Engert R, Walter T, Hancock REW. High-throughput generation of small antibacterial peptides with improved activity. *Nat. Biotechnol.* 2005; **23**: 1008–1012.
19. Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.* 1987; **30**: 1126–1135.
20. Jenssen H, Gutteberg TJ, Lejon T. Modelling the anti-herpes simplex virus activity of small cationic peptides using amino acid descriptors. *J. Pept. Res.* 2005; **66**: 48–56.
21. Hilpert K, Winkler DF, Hancock REW. 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. Protocol* 2007; **2**: 1333–1349.

DOI: 10.1002/psc