Genomics & Informatics Vol. 2(3) 121-125, September 2004

Application of Decision Tree for the Classification of Antimicrobial Peptide

Su Yeon Lee[1,2,†], Sunkyu Kim[2,†], Sukwon S. Kim[2], Seon Jeong Cha[1,2], Young Keun Kwon[3], Byung-Ro Moon[1,3,*] and Byeong Jae Lee[1,2*]

[1]Interdisciplinary Program in Bioinformatics, [2]Laboratory of Molecular Genetics, School of Biological Sciences, Institute of Molecular Biology and Genetics, and [3]Optimization Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea

Abstract

The purpose of this study was to investigate the use of decision tree for the classification of antimicrobial peptides. The classification was based on the activities of known antimicrobial peptides against common microbes including *Escherichia coli* and *Staphylococcus aureus*. A feature selection was employed to select an effective subset of features from available attribute sets. Sequential applications of decision tree with 17 nodes with 9 leaves and 13 nodes with 7 leaves provided the classification rates of 76.74% and 74.66% against *E. coli* and *S. aureus*, respectively. Angle subtended by positively charged face and the positive charge commonly gave higher accuracies in both *E. coli* and *S. aureus* datasets. In this study, we describe a successful application of decision tree that provides the understanding of the effects of physicochemical characteristics of peptides on bacterial membrane.

Keywords: decision tree, classification, antimicrobial peptides

Introduction

During the last two decades, a number of studies on bactericidal peptides have been carried out for the purpose of industrial, pharmaceutical and medical use with great interest (Park *et al.*, 1994; Kim *et al.*, 2003; Zasloff *et al.*, 2002). Naturally occurring antimicrobial peptides and a set of artificial peptides have been used as templates for generating synthetic peptides with substitutions and deletions of amino acids to analyze the interaction relationship between bacterial membrane and the physicochemical properties of peptides such as hydrophobicity, hydrophobic moment and net charge (Dathe *et al.*, 2001; Hong *et al.*, 2001; Dathe *et al.*, 2002; Dathe *et al.*, 1997). However, as demonstrated by numerous studies on structure-activity relationship (SAR), a precise interpretation of differences in the activity of antimicrobial peptides is often difficult due to the existence of complex interaction between the peptides and bacterial membrane (Rocca *et al.*, 1999a; Rocca *et al.*, 1999b).

A decision tree is a kind of supervised machine learning method that is often appropriate for describing complex nonlinear relationships between quantitative features and some a priori classification (both quantitative and qualitative) (Mitchell *et al.*, 1997). One advantage of the decision tree approach is that the decision rules can be displayed in easy-to-interpret graphical manner. Unlike many other machine-learning methods that rely on abstract interpretation of the feature space (Vapnik *et al.*, 1995), the decision tree gives sequential decision splits (nodes of the tree) based on some marginal feature sets, which in turn can be interpreted in terms of the biological or chemical process; in our case, in terms of peptide properties.

Materials and methods

Data set

A total of 402 peptides were collected from the literature of antimicrobial peptides to form a data set. It consists of natural antibiotic peptides and their analogs generated by changing physicochemical parameters of the natural antibiotic peptides. These peptides were divided into two groups according to their reported minimal inhibitory concentration (MIC) values that present the strength of antimicrobial activities in molar concentrations. Peptides with MIC values of 10 μM or below were grouped as strong peptides in terms of antimicrobial activity, whereas those with MIC values above 40 μM were grouped as weak antimicrobial peptides. Among the peptide, we used 133 antimicrobial peptides which had MIC values above 40 or below 10 μM against *E. coli*. We excluded data between 10 μM and 40 μM , because it is

*Corresponding author: E-mail imbglmg@plaza.snu.ac.kr(B. J. L.) moon@soar.snu.ac.kr (B-R. M.),
Tel +82-2-880-6775, Fax +82-2-872-9019
Accepted 19 July 2004
†These authors contributed equally to this work.

difficult to determine strong or weak. In the same way, we included 102 antimicrobial peptides with MIC values against *S. aureus*.

The data set of *E. coli* consists of 74 peptides with strong antimicrobial activity and 59 peptides with weak antimicrobial activity. In case of *S. aureus*, 41 peptides were strong and the rest were weak with respect to their antimicrobial activities.

Decision trees

Decision trees have been applied to various studies for biological data mining, such as protein secondary structure prediction (Selbig *et al.*, 2002), gene prediction in vertebrate DNA sequences (Salzberg *et al.*, 1998) and small molecule solubility prediction (Xia *et al.*, 2003). A classification and prediction of a given dataset by a decision tree is achieved by constructing a rooted tree graph with the leaves of the tree labeled with the values of the classification variable (e.g., strong/weak antibacterial activity) and the intermediate nodes each representing a test based on some feature subset. Each branch of the decision tree is labeled with the critical values for the test specified by the node above the branch (Quinlan *et al.*, 1993). In our application of a decision tree to the SAR of anti-microbial peptides, the physicochemical properties of peptides were the basis of tests at intermediate nodes and the resultant terminal leaves were labeled with the strength of antimicrobial activity. Structural parameters, such as peptide helicity, hydrophobicity, hydrophobic moment, peptide charge and the size of the hydrophobic/hydrophilic domain were influenced on membrane activity. Furthermore, it is assessed that the potential of these parameters increase antimicrobial activity (Dathe *et al.*, 1999). Of many physicochemical properties of peptides, the following attributes were selected as the feature set for the decision tree algorithm: Kyte-Doolittle hydrophobicity (Kyte *et al.*, 1982), maximum hydrophobicity (maxH) which is the best hydrophobic score obtained by calculating the average of hydrophobicity (H; window size was 5), and hydrophobic moment (Hm). Hydrophobic moment is the hydrophobicity of a peptide measured for a specified angle of rotation per residue. The strength of each periodic component is the quantity that has been termed the hydrophobic moment. For example, alpha helices tend to have strong periodicity in the hydrophobicity of about 3.6 residues and beta sheets about 2.3 residues. It means that many peptide sequences tend to form the periodic structure that maximizes their amphiphilicity. Amphiphilicity of the peptides has been reported as the most important factor governing antimicrobial activity compared to mean hydrophobicity or alpha-helix content (Pathak *et al.*, 1995). We also include other factors like helicity (He) predicted by NNPREDICT which is a web-based secondary structure program (Kneller *et al.*, 1990), angle subtended by positively charged face (A), net charges (netC), sum of positive charges (pC) and sum of negative charges (nC). Furthermore, additional attributes confined to the helical portion of the peptides (i.e. spatial arrangement of each amino acid) were selected to increase the accuracy of classification; they include hydrophobicity (H/h), net charges (netC/h), positive charges (pC/h), hydrophobic moment (Hm/h), angle subtended by positively charged face (A/h) and negative charges (nC/h) in helical region.

In this study, the decision was implemented by C4.5. C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan (Quinlan *et al.*, 1993). This program classifies antimicrobial peptides based on selected attributes which are treated as statistical property called information gain that measures how well a given attribute separates the training examples according to their target classification (Mitchell *et al.*, 1997). Information gain can be further described by entropy that characterizes the purity and/or complexity of an arbitrary collection of data. In the case of peptide data set (*S*), the entropy of set *S* relative to binary classification is defined as

$$Entropy(S) = -P_p \log_2 P_P - P_n \log_2 P_n \tag{1}$$

where P_P is the proportion of positive data (strong antimicrobial activity) in S and P_n is the proportion of negative data (weak antimicrobial activity) in S. (In this study, any entropy with 0log0 is defined to be 0.) Therefore, information gain can be expressed by the expected reduction in entropy caused by partitioning the data set according to selected attributes. More precisely, the information gain, *Gain* (*S, A*) of attribute *A*, relative to a collection of data set S is defined as

$$Gain(S, A) = Entropy(S) - \sum_{\nu \in V(A)} Entropy \frac{|S_\nu|}{|S|} (S_\nu) \tag{2}$$

where *V*(*A*) is the set of all possible values for attribute *A*, and *Sv* is the subset of *S* for which attribute *A* has value *v*. *Gain*(*S*, *A*) is, therefore, the expected reduction in entropy caused by the knowledge of the value of attribute *A* (Mitchell *et al.*, 1997).

Test all combinations of subsets of features

All combinations of 14 features (2^{14} = 16384) were tested. We obtained the results of optimal subset of features which the training error was smallest in each data set.

Evaluation of the classification

The binary classification produced by the decision tree was evaluated by using a stratified 5-fold cross-validation (Stone *et al.*, 1974; Weiss *et al.*, 1991). The training set is divided into 5 equal-sized subsets such that each example appears in exactly one test set. The folds were stratified so that they would contain the same proportions of classes as the original data set. For each subset, a decision tree is constructed from examples of the 4 other subsets and tested on examples from the excluded set. The average error rate over the 5 test sets is the error rate of a decision tree generated from all the data (Quinlan *et al.*, 1993).

A decision tree that is too complex may result in overfitting the data. As a result, this decision tree may classify the training data excellently but perform poorly on some test data. To avoid this problem, we employed a pruning method described in (Quinlan *et al.*, 1993) to make the decision tree simpler and more reasonable. Another approach to avoid this problem is to use a validation set. We divided a training data set into two. A half of a training set is used to form the decision tree using c4.5 and the other of training set is used as a validation set, which is used to generate rules from the decision tree using c4.5 rules (Quinlan *et al.*, 1993).

The overall rate of success with respect to classification accuracy is defined as

$$Accuracy=\frac{(TP+TN)}{(TN+FP+FN+TP)}\times 100 \tag{3}$$

where *TP* (true positive) is the number of correctly classified peptides with strong antimicrobial activity, *TN* (true negative) is the number of incorrectly classified peptides with weak antimicrobial activity, *FP* (false positive) is the number of incorrectly classified peptides with weak antimicrobial activity, and *FN* (false negative) is the number of incorrectly classified peptides with strong antimicrobial activity.

Several other statistics are useful for more detailed evaluation of the performance of the algorithms.

Sensitivity and selectivity are often used for better evaluation of the precision of algorithms. Sensitivity (S_n, true positive rate) represents the ability to detect positive instances, i.e. the proportion of correctly classified peptides with strong antimicrobial activity among all antimicrobial peptides with strong antimicrobial activity. It is a significant factor for determining the classification efficiency, and it is defined as

$$S_n=\frac{TP}{TP+FN} \tag{5}$$

Selectivity (S_e, true negative rate) describes the proportion of correctly classified peptides with low potency amongst all antimicrobial peptides with low potency. This is defined as:

$$S_e=\frac{TN}{FP+TN}\,. \tag{6}$$

Furthermore, in order to evaluate the validity of the rate of classification, an F-measure is obtained. F-measure is the sum of average of the measures of precision and recall. Precision value is a mathematical presentation of false-positives present in the final classification while the value for recall contains similar information as in S_n.

$$<ision=\frac{TP}{TP+FN} \tag{7}$$

$$recall=\frac{TN}{FP+TN} \tag{8}$$

Therefore, the F-measure can be formulated as below

$$\begin{aligned}F-measure&=\frac{1}{a\times\frac{1}{<ision}+(1-a)\times\frac{1}{recall}}\\&=\frac{2\times<ision\times recall}{<ision+recall}\end{aligned} \tag{9}$$

(Generally, $\alpha=0.5$)

Result

The decision tree for the classification of the peptide data set against *E. coli* was constructed by using subset of selected features. It composed of 17 nodes with 9 leaves. The rate of correct prediction for *E. coli* was approximately 76.64%. The sensitivity was 86.19% while the selectivity was 62.73%. Furthermore, 72.61% was achieved for the F-measure suggesting significant accuracy in the classification. The accuracy and a graphical illustration of the classification of the data set is shown in Fig. 1.

The overall classification accuracy was 74.66% for the data set classified with *S. aureus*. The tree (13 nodes with 7 leaves) was simpler than that obtained from the data set of *E. coli*. The sensitivity, selectivity and F-measure were 66.67%, 80.51% and 72.94%, respectively. The accuracy and F-measure were about the same in both microbes. The selectivity obtained from the data set of *S. aureus* (80.51%) was relatively higher than that from the data set of *E. coli* (62.73%). By contrast, the sensitivity of *E. coli* (86.19%) is higher than *S. aureus* (62.73%).

To compare with the effect of validation set, we constructed decision tree with and without validation set. As you see in table 1, the accuracy of using validation set was improved 5.25% and 3.98%, *E. coli, S. aureus*, respectively (Table 1).

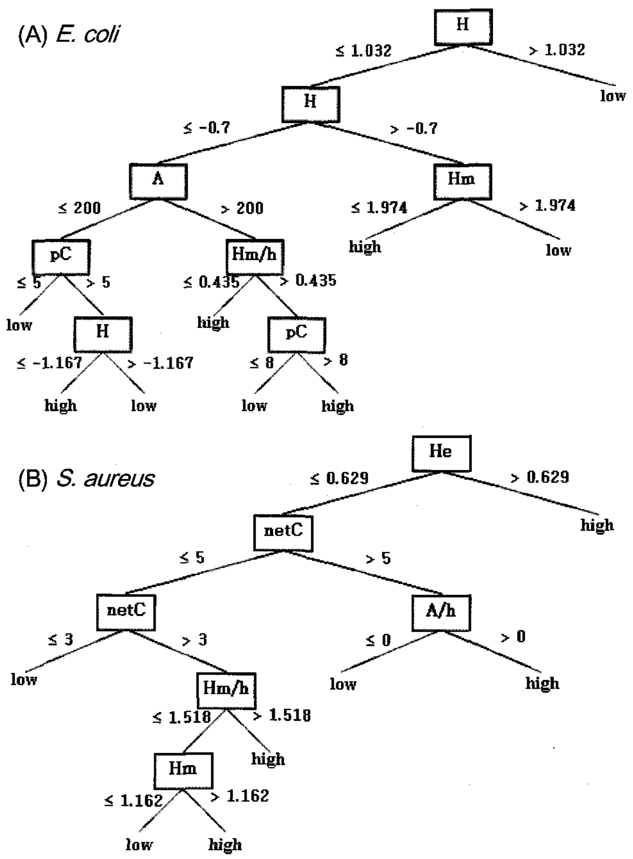

Fig. 1. A final pruned decision tree for *E. coli* (A) and *S. aureus* (B) showing the sequential decision tests for strong/weak anti-microbial action of peptides. Each node represents a classification test based on the indicated physico-chemical variable. The branches are labeled with critical values for the tests

Table. 1. The accuracy of general and our approach

microbes	without validation set (A)	with validation set (B)	improvement (B-A)
E. coli	71.50%	76.74%	5.24%
S. aureus	70.68%	74.66%	3.98%

Discussion

Classification is the process of finding a set of models that describe and distinguish data classes or concepts. Results in this study show that a decision tree with high-quality classification accuracy can be constructed for the classification of antimicrobial peptides vis-a-vis the strength of antimicrobial activity against *E. coli* and *S. aureus*. When comparing the rates of classification accuracy predicted for *E. coli* with that of *S. aureus*, the rate of classification accuracy for *E. coli* was approximately 2.08% higher than that of *S. aureus*. However, when considering the fact that the peptide data set for *S. aureus* was unevenly distributed between the strong and weak peptides, the classification result for *S. aureus* is as significant as the classification for *E. coli*. The values for F-measure are more suitable measures for estimating the performance for strong antimicrobial peptides, and the sensitivity(%) value of 86.19% for *E. coli* is higher than that for *S. aureus* (66.67%). Along this line, the classification accuracy may be further improved with evenly distributed data set.

As shown in figure 1, the attribute measuring positive charges in helix region (H) was chosen as the first intermediate node to classify in *E. coli*. Attributes selected by information gain were used to sort peptides further down the tree to obtain the rate of classification accuracy of 76.74%. It is evident from the tree that certain attributes were crucial in achieving high classification accuracy. Out of 14 attributes selected, three attributes (H, A and Hm) that were selected gave higher accuracies in *E. coli* data set and (He, netC and A/h) attributes were selected in *S. aureus* data sets. Selected features were varied according to bacteria. As mentioned previously, attributes are physicochemical properties of the antimicrobial peptides. The decision tree for the classification of antimicrobial peptides suggest that these features play crucial role in the strength of antimicrobial activity, and this finding concurs with previous reports (Dathe *et al.*, 1999).

The results obtained through this study indicate that a decision tree algorithm can be an efficient tool for predicting the activity of antimicrobial peptides. In particular, the algorithm allows us to extract important physico-chemical properties, which may be useful for rational peptide design for many therapeutic uses.

Acknowledgements

The authors thank Dr. Junhyong Kim for critical reading this manuscript.

References

Dathe, M., Wieprecht, T., Nikolenko, H., Handel, L., Maloy, W. L. , MacDonald, D.L., Beyermanna, M., and Bienerta M. (1997). Hydrophobicity, hydrophobic moment and angle subtended by charged residues modulate antibacterial and haemolytic activity of amphipathic helical peptides. *FEBS Lett.* 403, 208-212.

Dathe, M. and Wieprecht, T. (1999). Structural features of helical antimicrobial peptides: their potential to modulate activity on model membranes and biological cells. *Biochim. Biophys. Acta.* 1462, 71-87.

Dathe, M., Nikolenko, H., Meyer, J., Beyermann, M. and Bienert, M. (2001). Optimization of the antimicrobial activity of magainin peptides by modification of charge. *FEBS Lett.* 501, 146-150.

Dathe, M., Meyer, J., Beyermann, M., Maul, B., Hoischen, C., and Bienert, M. (2002). General aspects of peptide selectivity towards lipid bilayers and cell membranes studied by variation of the structural parameters of amphipathic helical model peptides. *Biochim. Biophys. Acta.* 1558, 171-186.

Hong, S.Y., Park, T.G., and Lee, K.H.. (2001). The effect of charge increase on the specificity and activity of a short antimicrobial peptide. *Peptides* 22, 1669-1674.

Kim, S., Kim, S.S., Bang, Y.J., Kim, S.J. and Lee B.J. (2003). In vitro activities of native and designed peptide antibiotics against drug sensitive and resistant tumor cell lines. *Peptides* 24, 945-953

Kneller, D.G,, Cohen, F.E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214, 171-182.

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105-132.

Mitchell, T.M. (1997). *Machine Learning* (The McGrow-Hill Companies Inc.)

Park, J.M., Jung, J.E., and Lee, B.J. (1994). Antimicrobial peptides from the skin of a Korean frog, *Rana rugosa. Biochem. Biophys, Res. Comm.* 205, 948-954

Pathak, N., Salas-Auvert, R., Ruche, G., Janna, M.H., McCarthy, D., and Harrison, R.G. (1995). Comparison of the effects of hydrophobicity, amphiphilicity, and alpha-helicity on the activities of antimicrobial peptides. *Proteins* 22, 182-186

Quinlan, J.R. (1993). *C4.5:Programs for Machine Learning* (Morgan Kaufmann Publisher Inc.)

Rocca, P.L., Biggin, P.C., Tieleman, D.P., and Sansom MSP. (1999a). Simulation studies of the interaction of antimicrobial peptides and lipid bilayers. *Biochim. Biophys. Acta.* 1462, 185-200.

Rocca, P.L., Shai, Y., and Sansom, M.S.P. (1999b). Peptide -bilayer interactions: simulations of dermaseptin B, an antimicrobial peptide. *Biophys. Chem.* 76, 145-159.

Salzberg, S., Delcher, A.L., Fasman K.H., and Henderson L. (1998). A decision tree system for finding genes in DNA. J. Comput. Biol. 5, 667-680.

Selbig, J., Mevissen, T., and Lenauer, T. (1999). Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* 15, 1039-1046.

Stark, M., Liu, L.P., and Deber, C.M. (2002). Cationic hydrophobic peptides with antimicrobial activity. *Antimicrob. Agents Chemother.* 46, 3585-3590.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Stat. Soc. B.* 36, 111-147.

Tachi, T., Epand, R.F., Epand, R.M., and Matsuzaki, M. (2002). Position-dependent hydrophobicity of the antimicrobial magainin peptide affects the mode of peptide-lipid interactions and selective toxicity. *Biochemistry* 41, 10723-10731.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory* (New York : Springer Verlag)

Weiss, S.M. and kulikowski, C.A. (1991). *Computer systems that learn*: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. (San Mateo : Morgan Kaufmann Publisher Inc.)

Xia, X., Maliski, E., Cheetham, J., and Poppe, L. (2003). Solubility prediction by recursive partitioning. *Pharm. Res.* 20, 1634-1640.

Zasloff, M. (2002). Antimicrobial peptides of multicellular organisms. *Nature* 415, 389-395.