Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Note

Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis

Wei Chen, Liaofu Luo [*]

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot, 010021, China

ARTICLE INFO

Article history:
Received 5 February 2009
Received in revised form 20 March 2009
Accepted 30 March 2009
Available online 5 April 2009

Keywords:
Antimicrobial peptides
Increment of diversity
Quadratic discrimination
Dipeptide composition

ABSTRACT

Accurate classification of antimicrobial peptides according to their biological activities will facilitate the design of novel antimicrobial agents and the discovery of new therapeutic targets. In this work, an excellent algorithm of Increment of Diversity with Quadratic Discriminant analysis (IDQD) was proposed to classify antimicrobial peptides with diverse biological activities.

Antimicrobial peptides, termed bacteriocins, are ribosomally synthesized short polypeptides, generally between 12 and 50 amino acids (Papagianni, 2003) and have been identified in various species ranging from prokaryotes to eukaryotes, such as bacteria, insects, plants, amphibians and mammals (Hancock, 2001; Hoffmann et al., 1996; García-Olmedo et al., 1998; Rinaldi, 2002; Lehrer and Ganz, 2002; Cole and Ganz, 2000). These peptides exhibit a potent activity against a wide range of microorganisms including gram-positive and gram-negative bacteria, protozoa, yeast, fungi, and even certain enveloped viruses and protozoa (Hancock and Lehrer, 1998; Toke, 2005; Tossi, 2005).

In contrast to many of the conventional antibiotics, antimicrobial peptides appear to be bacteriocidal instead of bacteriostatic and require a short contact time to induce the microbe-killing. In addition to these properties, they have been demonstrated to have a number of immunomodulatory functions that may be involved in the clearance of infection, including the ability to alter host gene expression, act as chemokines and/or induce chemokine production, inhibiting lipopolysaccharide induced pro-inflammatory cytokine production, promoting wound healing, and modulating the responses of dendritic cells and cells of the adaptive immune response.

Because bacteria may not easily acquire antibiotic resistance against them (Hancock and Patrzykat, 2002; Scott and Hancock, 2000; Bradshaw, 2003), antimicrobial peptides are excellent candidates for novel antimicrobial agents. Since their discovery in 1925 (Gartia, 1925), hundreds of antimicrobial peptides have been identified and some of them have been successfully used for defensing against both animal and human pathogens (Snelling, 2005; Kirkup, 2006). As well as having pathogen-lytic properties, antimicrobial peptides have also been shown to display anticancer/tumor activities (Kamysz et al., 2003). Thus, recognising the biological activities of antimicrobial peptides is critically important for the design of complements to conventional antibiotic therapy and novel therapeutic agents.

However, the classification of antimicrobial peptides is not yet explored and still at the infant stage. In this work, the improved method of Increment of Diversity with Quadratic Discriminant analysis (IDQD) was presented to classify the four categories of antimicrobial peptides (antiviral/HIV, anticancer/tumor, antibacterial and antifungal).

The 1228 mature antimicrobial peptides were extracted from the APD2 (Wang et al., 2009) database. The redundancy peptides were removed by using CD-HIT algorithm (Li et al., 2001). With the sequence identity less than 40%, we got 37 antiviral/HIV, 41 anticancer/tumor, 389 antibacterial, 177 antifungal peptides with single biological activity and 76 peptides with multiple biological activities (including 18 antiviral/HIV and antifungal, 38 antifungal and antibacterial, 20 antifungal and anticancer/tumor peptides, others were neglected due to their insufficient samples) according to the database annotation. These peptides were then divided into two datasets, training and testing. The training dataset contains the 37 antiviral/HIV, 41 anticancer/tumor peptides and 40 antibacterial and 40 antifungal peptides randomly selected from the 389 antibacterial and 177 antifungal peptides, respectively. The remaining 349 antibacterial, 137 antifungal peptides and 76 multiple biological activity peptides were designated as the independent test samples.

The IDQD method, suitable for the classification of two types of samples, was proposed and successfully applied in the prediction of exon–intron splice sites (Zhang and Luo, 2003). However, the classification of multiple types of samples is generally more difficult than that of two. Here, the improved form of the method was described, capable for the classification of multiple types of samples, and applied to the identification of antimicrobial peptides.

[*] Corresponding author. Tel.: +86 471 4992676; fax: +86 471 4993124.
E-mail addresses: lolfcm@mail.imu.edu.cn, chenwei_imu@yahoo.com.cn (L. Luo).

0167-7012/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2009.03.013

In the state space of d dimensions, the diversity measure of the diversity source S: $\{n_1, n_2, \ldots, n_d\}$ is defined as (Laxton, 1978),

$$D(S) = D(n_1, n_2, \cdots, n_d) = N\log_2 N - \sum_{i=1}^{d} n_i \log_2 n_i \quad N = \sum_{i=1}^{d} n_i \tag{1}$$

where n_i is the occurrence of the i-th character in diversity source S, and if n_i equals zero, then $n_i \log_2 n_i = 0$.

In the same dimensional space, for a query peptide X to be classified, the increment of diversity (ID) between the peptide X and the standard source S^α (α = antiviral/HIV, anticancer/tumor, antifungal and antibacterial peptides) can be defined as follows,

$$ID(X, S^\alpha) = D(X + S^\alpha) - D(X) - D(S^\alpha) \tag{2}$$

where $D(X)$ and $D(S^\alpha)$ are the diversity measure of the peptide X and standard source S^α, respectively. $D(X + S^\alpha)$ denotes the diversity measure of the mixed source $S^\alpha + X$. ID gives the relation of peptide X with standard source S^α. The smallest the ID, the most intimate relation of the inquired peptide X to standard source S^α.

When there are r sets of characters for a peptide X, we obtain an r-dimensional feature vector $\mathbf{R}(\mathrm{X}) = (ID_1, \ldots, ID_\mathrm{r})$ and need to integrate it into a nonlinear discriminant function ξ by quadratic discriminant (QD) analysis (Zhang, 1997), deduced from Bayes's theorem and under the assumption of ID_1 to ID_r obeying the r-dimensional normal distribution (Luo and Lv, 2007). The discriminant function ξ that gives the classification of the potential peptide X is expressed as follows (Zhang and Luo, 2003; Lv and Luo, 2008; Feng and Luo, 2008),

$$\xi_\alpha = \log_2 N_\alpha - \frac{(\mathbf{R} - \mu_\alpha)^T \sum_\alpha^{-1} (\mathbf{R} - \mu_\alpha)}{2} - \frac{1}{2}\log_2 |\Sigma_\alpha| \tag{3}$$

where μ_α is the average of $\mathbf{R}$ over all N_α samples (α = antiviral/HIV, anticancer/tumor, antibacterial and antifungal) in the training set and $\sum_\alpha$ is the corresponding $r \times r$ covariance matrix.

For the single biological activity peptide, the predictive result is made by assigning the query peptide X to the α-th class with which ξ_α has the maximum value,

$$\xi_\alpha = \mathrm{Max}\left\{\xi_{\mathrm{antiviral/HIV}}, \xi_{\mathrm{anticancer/tumor}}, \xi_{\mathrm{antibacterial}}, \xi_{\mathrm{antifungal}}\right\} \tag{4}$$

where the operator Max means taking the maximum value of ξ. For the multiple biological activity peptide, say two, if the second maximum ξ_β, say $\xi_{\mathrm{antibacterial}}$, is within a deviation of θ from the largest ξ_α, say $\xi_{\mathrm{antiviral/HIV}}$, i.e., $\xi_\alpha - \xi_\beta < \theta$, then the query peptide will be assigned to the β-th class as well. Here, $\theta = \mathrm{Min}\{|\xi_\alpha - \xi_\beta|, \alpha \neq \beta\}$ is the smallest deviation between the ξ values of the considered two classes (α and β) of single biological activity peptides. In addition, by considering the third or fourth maximum ξ value and complying with the analogous rules, we could give a classification for a query peptide with three or four biological activities.

According to the definition of Gromiha et al. (2005), we firstly calculated the compositions of the 20 amino acids and observed that the distribution of residues Cys, Gly, Val, Ile, Leu, Tyr and Trp show subtle difference among families. Further, the compositional differences of these residues are found to be statistically significant ($p < 0.001$) by using the analysis of variance (ANOVA). The residues Cys/Gly and Val are preferred by antiviral/HIV and anticancer/tumor peptides, while the strongly hydrophobic residues Ile/Leu and the weakly hydrophilic residues Tyr/Trp present the biasness to antibacterial and antifungal peptides, respectively. Thus, using the absolute occurrences of the 20 amino acids (AA), four IDs (ID_1, ID_2, ID_3 and ID_4) between peptide X and the four standard sources (antiviral/HIV, anticancer/tumor, antibacterial and antifungal peptides) are defined.

Since several works (Chou, 2005; Wang et al., 2006; Xiao et al., 2006a,b; Zhang et al., 2006; Zhou and Cai, 2006) have demonstrated that dipeptide composition can improve the prediction quality of protein classification, the other four IDs (ID_5, ID_6, ID_7 and ID_8) between peptide X and the four standard sources are deduced according to the absolute occurrences of the 400 dipeptides. Therefore, any query peptide can be represented by the 8-dimensional vector $\mathbf{R} = (ID_1, ID_2, \ldots, ID_8)$. Of the 8 IDs, the first four reflect the effect of AA composition and the last four reflect the effect of dipeptide, depicting the correlation of proximate residues.

A prediction method should be evaluated by the most rigorous and objective jackknife test (Chou and Zhang, 1995). In the jackknife test, each peptide in the training dataset is in turn singled out as an independent "test sample" and all the rule-parameters are calculated without using this one. Therefore, the jackknife test was applied to examine the performance of the IDQD method for the classification of antimicrobial peptides. The predictive results by using different parameters are listed in Table 1.

Though the antiviral/HIV, anticancer/tumor, antibacterial and antifungal peptides can be classified only by using the dipeptide composition, the predictive accuracies are still approximately 5% lower than that by using the parameters of AA composition and dipeptide

Table 1
Comparative results for the classification of antimicrobial peptides by the jackknife test*.

Parameters	Antimicrobial peptides	Sn (%)	Sp (%)	Acc (%)	MCC
AA (IDQD)	Antiviral/HIV	52.38	50.00	66.13	0.25
	Antifungal	44.44	35.29	52.56	0.01
	Anticancer/tumor	46.43	43.33	56.16	0.09
	Antibacterial	22.73	35.71	61.19	0.03
Dipeptide (IDQD)	Antiviral/HIV	80.95	80.95	90.70	0.75
	Antifungal	78.57	75.86	86.67	0.69
	Anticancer/tumor	77.78	80.77	86.67	0.68
	Antibacterial	81.82	81.82	90.70	0.76
AA and Dipeptide (IDQD)	Antiviral/HIV	**95.23**	**83.33**	**94.38**	**0.85**
	Antifungal	**84.07**	**90.91**	**89.36**	**0.76**
	Anticancer/tumor	**89.29**	**83.33**	**90.32**	**0.80**
	Antibacterial	**90.91**	**90.48**	**95.46**	**0.85**
AA and Dipeptide (SVM)	Antiviral/HIV	71.43	75.00	90.58	0.65
	Antifungal	81.48	78.57	87.50	0.71
	Anticancer/tumor	78.57	75.86	86.52	0.67
	Antibacterial	81.82	85.71	91.67	0.78

*Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthew's correlation coefficient.

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},\ \mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}},\ \mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}$$

TP, true positive; FN, false negative; TN, true negative; FP, false positive.

composition together (Table 1). These results imply that the dipeptide composition plays a major role for the classification of antimicrobial peptides, while the information extracted from the AA composition is also an extremely important complements and can significantly improve the prediction quality (Table 1). Consequently, all the following predictions are carried out by using the combination of these two kinds of parameters (AA and dipeptide composition).

In addition, we compared the predictive capability of our IDQD method with that of the support vector machine (SVM, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm) for the classification of antimicrobial peptides in the same training dataset. The 8-dimensional feature vector, same as that for IDQD, was used as the input into the SVM classifier. The best predictive result of SVM was achieved using the radial basis function (RBF) kernel, shown in the last four rows of Table 1. We found that the successful rates of our IDQD model are superior to that of the SVM model.

To roundly assess its reliability, our IDQD model was also validated on the independent dataset. Our proposed algorithm correctly identified the antibacterial and antifungal peptides in the independent dataset with the sensitivities of 87.11% and 83.94%, respectively. Additionally, our model also accurately picked out the three types of multiple biological activity peptides deposited in the test dataset with the sensitivities of 73.42%, 70.17% and 75.85%, respectively. These results further emphasized that the improved IDQD method is a promising approach for the classification of antimicrobial peptides.

In conclusion we have shown that the improved IDQD method can be applied to the classification of antimicrobial peptides with one or more biological activities. The efficient extraction of sequence information by use of diversity measure in the high-dimensional space and the synthesis of different types of sequence information into one discriminant function ξ are two important factors for the success of IDQD algorithm.

The successful predictive results demonstrate that antimicrobial peptides can be accurately classified by integrating the information deposited in AA and dipeptide, and further indicate that the different residue compositions among antiviral/HIV, anticancer/tumor, antibacterial and antifungal peptides may be the cause of their biological activities. The prediction accuracy may be further improved with future accumulation of knowledge regarding antimicrobial peptides, peptide sample collections in the dataset and more physicochemical properties about the peptides. We hope that the improved IDQD model proposed here will be an effective tool for the classification of antimicrobial peptides.

Acknowledgments

We wish to express our gratitude to the Editors and two anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this article. This work was supported by National Natural Science Foundation of China (No. 90403010).

References

Bradshaw, J.P., 2003. Cationic antimicrobial peptides: issues for potential clinical use. BioDrugs 17, 233–240.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Cole, A.M., Ganz, T., 2000. Human antimicrobial peptides: analysis and application. Biotechniques 29, 822–831.

Feng, Y.E., Luo, L.F., 2008. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35, 607–614.

García-Olmedo, F., Molina, A., Alamillo, J.M., Rodríguez-Palenzuéla, P., 1998. Plant defense peptides. Biopolymers 47, 479–491.

Gartia, A., 1925. Sur un remarquable exemple d'antagonisme entre deux souches de colibacille. Comput. Rend. Soc. Biol. 93, 1040–1041.

Gromiha, M.M., Ahmad, S., Suwa, M., 2005. Application of residue distribution along the sequence for discriminating outer membrane proteins. Comput. Biol. Chem. 29, 135–142.

Hancock, R.E., 2001. Cationic peptides: effectors in innate immunity and novel antimicrobials. Lancet, Infect. Dis. 1, 156–164.

Hancock, R.E., Lehrer, R., 1998. Cationic peptides: a new source of antibiotics. Trends Biotechnol. 16, 82–88.

Hancock, R.E., Patrzykat, A., 2002. Clinical development of cationic antimicrobial peptides: from natural to novel antibiotics. Curr. Drug Targets. Infect. Disord. 2, 79–83.

Hoffmann, J.A., Reichhart, J.M., Hetru, C., 1996. Innate immunity in higher insects. Curr. Opin. Immunol. 8, 8–13.

Kamysz, W., Okruèj, M., Łukasiak, J., 2003. Novel properties of antimicrobial peptides. Acta Biochim. Pol. 50, 461–469.

Kirkup Jr., B.C., 2006. Bacteriocins as oral and gastrointestinal antibiotics: theoretical considerations, applied research, and practical applications. Curr. Med. Chem. 13, 3335–3350.

Laxton, R.R., 1978. The measure of diversity. J. Theor. Biol. 70, 51–67.

Lehrer, R.I., Ganz, T., 2002. Defensins of vertebrate animals. Curr. Opin. Immunol. 14, 96–102.

Li, W.Z., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 17, 282–283.

Luo, L.F., Lv, J., 2007. Sequence pattern recognition in genome analysis. Computation in Modern Science and Engineering: Proceedings of the International Conference on Computational Methods in Science and Engineering 2007, vol. 963, pp. 1278–1281.

Lv, J., Luo, L.F., 2008. Prediction for human transcription start site using diversity measure with quadratic discrimination. Bioinformation 2, 316–321.

Papagianni, M., 2003. Ribosomally synthesized peptides with antimicrobial properties: biosynthesis, structure, function, and applications. Biotechnol. Adv. 21, 465–499.

Rinaldi, A.C., 2002. Antimicrobial peptides from amphibian skin: an expanding scenario. Curr. Opin. Chem. Biol. 6, 799–804.

Scott, M.G., Hancock, R.E., 2000. Cationic antimicrobial peptides and their multifunctional role in the immune system. Crit. Rev. Immunol. 20, 407–431.

Snelling, A.M., 2005. Effects of probiotics on the gastrointestinal tract. Curr. Opin. Infect. Dis. 18, 420–426.

Toke, O., 2005. Antimicrobial peptides: new candidates in the fight against bacterial infections. Biopolymers 80, 717–735.

Tossi, A., 2005. Host defense peptides: roles and applications. Curr. Protein Pept. Sci. 6, 1–3.

Wang, S.Q., Yang, J., Chou, K.C., 2006. Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J. Theor. Biol. 242, 941–946.

Wang, G., Li, X., Wang, Z., 2009. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res. 37, D933–D937.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006a. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30, 49–54.

Xiao, X., Shao, S.H., Huang, Z.D., Chou, K.C., 2006b. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J. Comput. Chem. 27, 478–482.

Zhang, M.Q., 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc. Natl. Acad. Sci. U. S. A. 94, 565–568.

Zhang, L.R., Luo, L.F., 2003. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res. 31, 6214–6220.

Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., Shi, J.Y., 2006. Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30, 461–468.

Zhou, G.P., Cai, Y.D., 2006. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 63, 681–684.