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 (Cartia, 1925), hundreds of antimicrobial peptides have been identified and some of them have been successfully used for defending against both animal and human pathogens (Snelling, 2005; Kirkup, 2006). As well as having pathogen-lytic properties, antimicrobial peptides also demonstrate the ability to display anticancer/tumor activities (Kamysz et al., 2003). Thus, recognizing 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 developed 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 73 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.
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),

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

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 query peptide and the standard source $S^α$ ($α = \text{antiviral/HIV, anticancer/tumor, antibacterial and antibacterial peptides}$) can be defined as follows,

$$ID(X, S^α) = D(X + S^α) - D(X) - D(S^α)$$

where $D(X)$ and $D(S^α)$ are the diversity measure of the peptide $X$ and standard source $S^α$, respectively. $D(X + S^α)$ denotes the diversity measure of the mixed source $S^α + X$. $ID$ 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 $R(X) = (ID_1, \ldots, ID_r)$ 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),

$$ξ_α = \log_2 N_α - \frac{(R - μ_α)\sum_{β=1}^{α} R - μ_β)}{2} - \frac{1}{2} \log_2 |Σ_α|$$

where $μ_α$ is the average of $R$ over all $N_α$ samples ($α = \text{antiviral/HIV, anticancer/tumor, antibacterial and antibacterial peptides}$) in the training set and $Σ_α$ 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,

$$ξ_α = \text{Max} \{ξ_{\text{antiviral/HIV}}, ξ_{\text{anticancer/tumor}}, ξ_{\text{antibacterial}}, ξ_{\text{antibacterial}}\}$$

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 $ξ_{\text{antibacterial}}$, is within a deviation of $θ$ from the largest $ξ_α$, say $ξ_{\text{antiviral/HIV}}$, i.e., $\xi_α - θ < ξ_β < θ$, then the query peptide will be assigned to the $β$-th class as well. Here, $θ = \min(|ξ_α - ξ_β|, ξ_β - θ) = \text{the smallest deviation between the ξ values of the considered two classes (α and β)}$ 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 hydrophobic 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 ($\text{antiviral/HIV, anticancer/tumor, antibacterial and antifungal peptides}$) 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 dipptide 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 dipptides. Therefore, any query peptide can be represented by the 8-dimensional vector $R = (ID_1, ID_2, \ldots, ID_8)$. Of the 8 IDs, the first four reflect the effect of AA composition and the last four reflect the effect of dipptide, 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.

Table 1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Antimicrobial peptides</th>
<th>Sn (%)</th>
<th>Sp (%)</th>
<th>Acc (%)</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA (IDQD)</td>
<td>Antiviral/HIV</td>
<td>52.38</td>
<td>50.00</td>
<td>66.13</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Antifungal</td>
<td>44.44</td>
<td>35.29</td>
<td>52.56</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Anticancer/tumor</td>
<td>46.43</td>
<td>43.33</td>
<td>56.10</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Antibacterial</td>
<td>22.73</td>
<td>35.71</td>
<td>61.19</td>
<td>0.03</td>
</tr>
<tr>
<td>Dipeptide (IDQD)</td>
<td>Antiviral/HIV</td>
<td>80.95</td>
<td>80.55</td>
<td>90.70</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Antifungal</td>
<td>78.57</td>
<td>75.86</td>
<td>86.67</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Anticancer/tumor</td>
<td>77.78</td>
<td>80.77</td>
<td>86.67</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Antibacterial</td>
<td>81.82</td>
<td>81.82</td>
<td>90.70</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>AA and Dipeptide (IDQD)</td>
<td>95.23</td>
<td>83.33</td>
<td>94.38</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Antiviral/HIV</td>
<td>84.07</td>
<td>90.91</td>
<td>89.36</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Antifungal</td>
<td>89.29</td>
<td>83.33</td>
<td>90.32</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Anticancer/tumor</td>
<td>90.91</td>
<td>90.48</td>
<td>95.46</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Antibacterial</td>
<td>71.43</td>
<td>75.00</td>
<td>90.58</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AA and Dipeptide (SVM)</td>
<td>81.48</td>
<td>78.57</td>
<td>87.50</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Antiviral/HIV</td>
<td>78.57</td>
<td>75.86</td>
<td>86.52</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Anticancer/tumor</td>
<td>81.82</td>
<td>85.71</td>
<td>91.67</td>
<td>0.78</td>
</tr>
</tbody>
</table>

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

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

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 complement 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 Σ 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


