Sequence analysis

Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types

Weizhong Lin¹,² and Dong Xu²,*

¹Information Engineering School, Jingdezhen Ceramic Institute, Jingdezhen 333406, China and ²Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 28, 2016; revised on August 7, 2016; accepted on August 22, 2016

Abstract

Motivation: With the rapid increase of infection resistance to antibiotics, it is urgent to find novel infection therapeutics. In recent years, antimicrobial peptides (AMPs) have been utilized as potential alternatives for infection therapeutics. AMPs are key components of the innate immune system and can protect the host from various pathogenic bacteria. Identifying AMPs and their functional types has led to many studies, and various predictors using machine learning have been developed. However, there is room for improvement; in particular, no predictor takes into account the lack of balance among different functional AMPs.

Results: In this paper, a new synthetic minority over-sampling technique on imbalanced and multi-label datasets, referred to as ML-SMOTE, was designed for processing and identifying AMPs’ functional families. A novel multi-label classifier, MLAMP, was also developed using ML-SMOTE and grey pseudo amino acid composition. The classifier obtained 0.4846 subset accuracy and 0.16 hamming loss.

Availability and Implementation: A user-friendly web-server for MLAMP was established at http://www.jci-bioinfo.cn/MLAMP.

Contacts: linweizhong@jci.edu.cn or xudong@missouri.edu

1 Introduction

With rapid increase in the infection resistance of antibiotics, it is urgent to find novel infection therapeutics. Over the past decade antimicrobial peptides (AMPs) have been utilized as potential alternatives for fighting infectious diseases. AMPs are key components of the innate immune system and can protect the host from various pathogenic bacteria. In invertebrates and vertebrates, AMPs have dual roles: rapid microbial killing and subsequent immune modulation (Wang, 2014). These effects result from AMP inducing multiple damages in bacteria by disrupting bacteria membranes (Malmsten, 2014), by inhibiting proteins, DNA and RNA synthesis, or by interacting with certain intracellular targets (Bahar and Ren, 2013). Therefore, AMPs were developed increasingly for new drugs. Some examples of using AMPs in therapeutics have been reported. Popovic et al. (2012) found that peptides with antimicrobial and anti-inflammatory activities had therapeutic potential for treatment of acne vulgaris. Yancheva et al. (2012) synthesized a novel dipeptide with antimicrobial activity against four of five tested bacterial strains of Escherichia coli. Conlon et al. demonstrated that peptides with antimicrobial activity from frog skin could stimulate insulin release, and hence had potential as an incretin-based therapy for Type 2 diabetes mellitus (Conlon et al., 2014). In addition,
AMPs have been used as anticancer peptides in cancer therapy (Gaspar et al., 2013).

A surge in research on AMPs has promoted the development of various databases and prediction tools. APD2 (Wang et al., 2009) is a system dedicated to establishing a glossary, nomenclature, classification, information search, prediction, design, and statistics of AMPs. It gathered 2544 AMPs from the literature. CAMP (Thomas et al., 2010; Waghu et al., 2013) holds 6756 antimicrobial sequences and 682 3D structures of AMPs, together with prediction and sequence analysis tools. Niarcho et al. (2013) tested all subsequences ranging from 5 to 100 amino acids of the plant proteins in UniProKB/Swiss-prot and constructed an AMP database for plant species, named C-PamP. Zhao et al. (2013) developed LAMP, a database used to aid the discovery and design of AMPs as new antimicrobial agents. The database contains 3904 natural AMPs and 1643 synthetic peptides. DBAASP was a manually curated database built by Gogoladze et al. (2014), and it collected those peptides for which antimicrobial activities against particular targets have been evaluated experimentally.

Generally, AMPs are short peptides with 10–50 amino acids (Malmsten, 2014) and have very low sequence homology to one another. So it is challenging to identify AMPs and its activities by automatic tools. Researchers made considerable efforts in this regard. In these studies, the support vector machine (SVM) was usually used as existing classifier learning algorithms to strengthen learning with respect to the imbalance problem in MLC. He et al. (2012) took into account the imbalance in predicting subcellular localization of human proteins. Charte et al. (2015) built an undersampling and oversampling algorithm on MLDs. Those studies improved the multi-label classification performance; however, they have some drawbacks in how to address the multi-label character of the new synthetic instance. In this paper, we tackle the imbalanced problem by a novel oversampling model referred to as ML-SMOTE, which is a synthetic minority oversampling on MLDs. We developed a new tool as a two-level AMP predictor based on ML-SMOTE. For a peptides sequence, we first identify whether it is an AMP. If yes, we then predict what potential activities it has. The first-level is a binary predictor, and the second-level predictor is an unbalanced and multi-labeled multi-classes predictor. The result shows ML-SMOTE can adjust the label set distribution to improve the performance of the predictor.

Among them, SMOTE (Synthetic Minority Oversampling Technique) (Chawla, 2010) is a state-of-art oversampling methods. Chawla (2010) argued that SMOTE creates effective regions for learning the minority class rather than being subsumed by the majority class samples around them. In bioinformatics, some studies have applied SMOTE to balance the skewing benchmark datasets (Jia et al., 2016; Li et al., 2015; Xiao et al., 2015). In addition, similar approaches have been recently introduced to handle the imbalanced datasets, such as Monte Carlo sampling (Jia et al., 2016c) and fusion ensemble approach (Liu et al., 2016a; Qiu et al., 2016).

Although the aforementioned methods have some success in addressing unbalanced datasets, they have not achieved a satisfactory result in processing multi-labeled and imbalanced datasets simultaneously. Few works address the imbalance problem in MLC. He et al. (2012) took into account the imbalance in predicting subcellular localization of human proteins. Charte et al. (2015) built an undersampling and oversampling algorithm on MLDs. Those studies improved the multi-label classification performance; however, they have some drawbacks in how to address the multi-label character of the new synthetic instance. In this paper, we tackle the imbalanced problem by a novel oversampling model referred to as ML-SMOTE, which is a synthetic minority oversampling on MLDs. We developed a new tool as a two-level AMP predictor based on ML-SMOTE. For a peptides sequence, we first identify whether it is an AMP. If yes, we then predict what potential activities it has. The first-level is a binary predictor, and the second-level predictor is an unbalanced and multi-labeled multi-classes predictor. The result shows ML-SMOTE can adjust the label set distribution to improve the performance of the predictor.
2 Methods

2.1 Benchmark dataset
The benchmark dataset \( S^{\text{Benchmark}} \) used in this study was taken from Xiao et al. (2013). The dataset can be formulated as
\[
S^{\text{Benchmark}} = S^+ \cup S^-
\]
where \( S^+ \) contains 879 AMPs, and \( S^- \) contains 2405 non-AMPs. The 879 AMPs are formulated as
\[
S^+ = \bigcup_{i=1}^{7} S_i
\]
where \( S_1 \) contains 770 antibacterial peptides, \( S_2 \) 140 anti-cancer/tumor peptides, \( S_3 \) 366 antifungal peptides, \( S_4 \) 84 anti-HIV peptides and \( S_5 \) 124 antiviral peptides.

2.2 Sequence encoding scheme
To develop a powerful method for classifying AMPs and their functional families according to the sequence information, one of the keys is to formulate the peptides with an effective mathematical expression that can truly reflect the intrinsic correlation with the target to be identified. However, when comparing with other protein functional predictions, the challenge is identifying how AMPs deal with shorter peptides. For a peptides sample \( P \) of \( L \) amino acids
\[
P = R_1 R_2 R_3 \ldots R_L
\]
where \( R_i \) (\( 1 \leq i \leq L \)) represents the \( i \)th residue, \( L \) is usually between 5 and 50.

In this study, we formulated an amino acids sequence by using Chou’s PseAAC (Chou, 2001, 2005) with the grey model (GM) (Deng, 1989). According to Chou’s general PseAAC formula (Chou, 2009, 2011), the peptides \( P \) in Eq. 3 can be represented as
\[
P = [p_1 \ p_2 \ \cdots \ p_k \ \cdots \ p_{20}]^T
\]
where \( T \) is a transpose operator, while the subscript \( \Omega \) is an integer and its value as well as the components \( p_1, p_2, \ldots \) depend on how to extract the desired information from the amino acid sequence of \( P \).

In our study, we use the GM(1,1) model, which is an important and generally used model in GM. GM(1,1) firstly converts a series without any obvious regularity into a strict monotonic increasing series by using the accumulative generation operation (AGO). This process can reduce the randomness and enhance the smoothness of the series and minimize any interference from the random information. Let us assume that
\[
X^{(0)} = \left( x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n) \right)
\]
is a non-negative original series of real numbers with an irregular distribution. Then
\[
X^{(1)} = \left( x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n) \right)
\]
is viewed as the first-order accumulative generation operation (1-AGO) series for \( X^{(0)} \), and the components in \( X^{(1)} \) are given by
\[
x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \quad k = 1, 2, \ldots, n
\]
The GM(1,1) model can be expressed by the following grey differential equation with one variable:
\[
\frac{dX^{(1)}}{dt} + aX^{(1)} = b
\]
where \( a \) and \( b \) are elements of parameters vector \( \hat{a} \), that is
\[
\hat{a} = [a, \ b]^T
\]
In Eq. 8, \( -a \) is the developing coefficient and \( b \) the influence coefficient. They can be solved using a least square estimator.
\[
\hat{a} = [a, \ b]^T = [B^T B]^{-1} B^T Y
\]
where
\[
B = \begin{bmatrix}
-0.5(x^{(1)}(1) + x^{(1)}(2)) & 1 \\
-0.5(x^{(1)}(2) + x^{(1)}(3)) & 1 \\
\vdots & \vdots \\
-0.5(x^{(1)}(n - 1) + x^{(1)}(n)) & 1 \\
\end{bmatrix}
\]
\[
Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}
\]
The coefficients \( a \) and \( b \) should carry some intrinsic information contained in the discrete data sequence \( X^{(0)} \) sampled from the system investigated. In view of this, we incorporate these coefficients into the general form of PseAAC (Eq. 4) to reflect the correlation between the peptide sequence and prediction labels. In order to translate an amino acid sequence expressed with alphabets in Eq. 3 into a non-negative real series in Eq. 5, we need the amino acid numerical codes. In the same manner as that shown in (Xiao et al., 2013), we also use the numerical value of the following five physical-chemical properties for each of the 20 amino acids: (1) hydrophobicity; (2) pk1 (\( \text{C}^\alpha - \text{COOH} \)); (3) pk2 (\( \text{NH}_3 \)); (4) PI (25°C); and (5) molecular weight. Finally, we used a 30-D features vector to represent a peptide; i.e. instead of Eq. 4, we now have
\[
P = [p_1, p_2, \ldots, p_{20}, p_{21}, \ldots, p_{30}]^T
\]
where \( p_i \) (\( 1 \leq i \leq 20 \)) are the frequencies of 20 amino acids; and \( p_{21} \) and \( p_{22} \) are the coefficients of Eq. 10 when amino acids are coded by hydrophobicity numerical values; \( p_{23} \) and \( p_{24} \) are the coefficients of Eq. 10 when amino acids are coded by \( \text{pK}1 \) numerical values, and so on.

2.3 ML-SMOTE algorithm
In Eq. 2, the AMP function family dataset is an unbalanced MLD, in which the antibacterial peptides have nine times the amount of the anti-HIV peptides. How to handle the MLC in unbalanced MLD is essential for improving prediction performance.

Let \( X \subset R^m \) denote an m-dimensions real vector of instance and let
\[
Y = \{ l_1, l_2, \ldots, l_q \}
\]
be a class label set. MLD can be represented as
\[
D = \{(x, y) \mid x \in X, \ y \subseteq Y \}
\]
We define the sample set with the \( j \)-th (\( 1 \leq j \leq q \)) label as
\[
D^{(j)} = \left\{ (x^{(0)}, y^{(0)}) \mid (x^{(0)}, y^{(0)}) \in D \text{ and } l_j \in y^{(0)} \right\}
\]
Algorithm ML-SMOTE algorithm’s pseudo-code

Inputs: Dataset: D with m features and q labels (see Eq. 17); 
k (the number of nearest neighbors)
Outputs: Preprocessed dataset S
(1) $S = D$
(2) MeanSize = $\frac{1}{2} \sum_{i=1}^{n} \|D^{(i)}\|$ (Ti is defined as Eq. 18)
(3) For $k = 1$ to $q$
(4) If $\|D^{(k)}\| < $MeanSize
(5) For each sample $t_i^{(k)}$ in $D^{(k)}$
(6) Find $k$-nearest neighbors set $knn$ of $t_i^{(k)}$
(7) Randomly select a sample $z$ from $knn$
(8) Get a random vector $(r_1, r_2, \ldots, r_m)$, where each element of $r$ is a random number between 0 and 1.
(9) Calculate features of new sample: $v = (1 - r_1) \cdot s_i + r_1 \cdot z_y$
(10) Add new sample $(v, u)$ to $S$
(11) End for
(12) End if
(13) End for

If $\|D_k\| \gg \|D_{l_i}\|$, the class $l_i$ is a majority class and the class $l_j$ is a minority class.

Different from SMOTE (Chawla et al., 2002) in a single label dataset, the new synthetic instance may have one or more labels. Hence, in (Charte et al., 2015), Charte et al. compared random undersampling (RUS) and random oversampling (ROS) based on Label Power-set (LP) and Multi-Label (ML), respectively. However, their LP-RUS and LP-ROS methods can only work well when the label density is low. Moreover, because their ML-ROS just clones the minority class samples, it is ineffective when these samples simultaneously have the majority class label, which happens often in MLD. In this study, we propose a novel oversampling model named ML-SMOTE. In the following algorithm description, we express a multi-label dataset (see Eq. 15) with N samples as

$$D = \{t_i = (x_i, y_i) | x_i = (x_{i1}, \ldots, x_{im}), y_i = (y_{i1}, \ldots, y_{iq}) | 1 \leq i \leq N \}$$  

where $y_{ij}$ = $\begin{cases} 1, & \text{if } x_i \text{ has } l_j \text{ label} \\ 0, & \text{otherwise} \end{cases}$ (1 \leq j \leq q)$

and the subset $D^{(k)}$ in which each sample is labeled $l_i$ class:

$$D^{(k)} = \{t_i^{(k)} = (x_i^{(k)}, y_i^{(k)}) | x_i^{(k)} = (x_{i1}, \ldots, x_{im}), y_i^{(k)} = (y_{i1}^{(k)}, \ldots, y_{iq}^{(k)}) | (1 \leq i \leq q) \}$$

For a new sample $(v, u)$ synthesized from $t_i^{(k)}$ and its near neighbor $z$ in $D^{(k)}$,

$$u_w = \begin{cases} 1 & \text{if } y_{iw}^{(k)} = 1 \text{ and } z_{yw} = 1 \\ 0 & \text{if } y_{iw}^{(k)} = 0 \text{ and } z_{yw} = 0 \\ 0 \text{ or } 1, & \text{randomly if } y_{iw}^{(k)} \neq z_{yw} \end{cases}$$

3 Results

After the sequence feature retrieval and ML-SMOTE preprocessing as described above, a two-level AMP predictor named MLAMP was constructed, in which the Ensemble of Classifier Chains (ECC) algorithm (Wagh et al., 2014) was adopted as the prediction method (Fig. 1). We used the canonical implementation of ECC provided by the MULAN (Tsoumakas et al., 2010; Tsoumakas et al., 2011) multi-label learning in the Weka (Hall et al., 2009) library. And for ECC, the binary and multi-class learners are implemented on the Weka platform using the Random Forest (RF) algorithm (Breiman, 2001).

MLAMP is a two-level prediction engine (See Fig. 1). The first level of MLAMP predicts a query peptide as AMP or non-AMP by using the RF algorithm. It belongs to the case of single-label classification. The following four measures were used for examining the performance of a single-label predictor, they are: (i) overall accuracy or Acc; (ii) Mathew’s correlation coefficient or MCC; (iii) sensitivity or Sn; and (iv) specificity or Sp.

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

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative.
Although Eq. 20 was often used in the literature to measure the prediction quality of a method, they often lack intuitiveness, especially to biologists, particularly the MCC. According to Chou’s formulation, these four measures can be expressed as (Chen et al., 2016; Lin et al., 2014)

\[
\begin{align*}
Sn &= 1 - \frac{N^-_c}{N^-} \quad 0 \leq Sn \leq 1 \\
Sp &= 1 - \frac{N^+_c}{N^+} \quad 0 \leq Sp \leq 1 \\
Acc &= 1 - \frac{N^+_c + N^-_c}{N^+ + N^-} \quad 0 \leq Acc \leq 1 \\
MCC &= 1 - \sqrt{\left(\frac{N^+_c - N^-_c}{N^+ + N^-}\right)^2 + \left(\frac{N^-_c - N^+_c}{N^- + N^+}\right)^2} \quad 1 \leq MCC \leq 1
\end{align*}
\]

where \(N^+_c\) stands for the total number of AMP samples investigated, whereas \(N^-_c\) for the number of AMP samples incorrectly predicted to be of non-AMP; \(N^+\) for the total number of non-AMP samples investigated, whereas \(N^-\) for the number of non-AMP samples incorrectly predicted to be of AMP. With such a formulation as given in Eq. 21, the meanings of sensitivity, specificity, overall accuracy and Mathew’s correlation coefficient and their scopes would become more intuitive and easier-to-understand, particularly for the Mathew’s correlation coefficient, as concurred by a series of studies published very recently (Jia et al., 2016; Tsoumakas et al., 2007, 2013; Tsoumakas and Katakis, 2007, 2010; Xiao et al., 2014a,b,c; Qiu et al., 2016b, 2016; Tsoumakas and Katakis, 2007; Tsoumakas et al., 2013). Suppose \(L_k\) is the subset that contains all the labels for the kth sample \(P_k\); \(L'_k\) is the subset that contains all the predicted labels for the kth sample \(P_k\); \(N\) is the total number of samples; and \(M\) is the total number of labels. In this study, \(N = 879\) and \(M = 5\). The five metrics have been clearly defined as follows (Chou, 2013):

\[
\begin{align*}
\text{Precision} &= \frac{1}{N} \sum_{k=1}^{N} \frac{||L_k \cap L'_k||}{||L_k||} \\
\text{Recall} &= \frac{1}{N} \sum_{k=1}^{N} \frac{||L_k \cap L'_k||}{||L'_k||} \\
\text{Accuracy} &= \frac{1}{N} \sum_{k=1}^{N} \frac{||L_k \cap L'_k||}{||L_k||} \\
\text{Subset Accuracy} &= \frac{1}{N} \sum_{k=1}^{N} \frac{\Delta(L_k, L'_k)}{M} \\
\text{Hamming Loss} &= \frac{1}{N} \sum_{k=1}^{N} \frac{\left(||L_k \cup L'_k|| - ||L_k \cap L'_k||\right)}{M}
\end{align*}
\]

where \(\|\|\) is the operator acting on the set therein to count the number of its elements, and

\[
\Delta(L_k, L'_k) = \begin{cases} 
1, & \text{if all the labels in } L_k \text{ are identical to those in } L'_k \\
0, & \text{otherwise}
\end{cases}
\]

When assessing a predictor, the following three cross-validation methods are often used in the literature: independent dataset test, subsampling (K-fold cross-validation) test and jackknife test. However, as elaborated in (Chou and Zhang, 1995), among the three cross-validation methods, the jackknife test is deemed the least arbitrary and most objective because it can always yield a unique result for a given benchmark dataset. Hence, the jackknife test was adopted in this study to examine the anticipated success rates of the current predictor. The process of jackknife test can be explained as follows:

\[
\text{Input: multi-label dataset } T = \{P_i \mid 1 \leq i \leq N\}. \\
\text{Output: predicted label set.} \\
\text{For } i: 0 \rightarrow N \\
\text{ } T \text{ is divided into testing dataset } T_s = \{P_i\}, \\
\text{and training dataset } T_r = T - T_s. \\
\text{Generate new training dataset } T' \text{ by using ML-SOMTE on } T_r. \\
\text{Train model on } T' \text{ by using ECC algorithm.} \\
\text{Predict the label set of } P_i \text{ by the model trained above.} \\
\text{End For}
\]

Table 1 compares the performance of MLAMP with an existing method iAMP-2L in the first-level result on the benchmark \(g^{Bench}\) (Eq. 1), where overall accuracy Acc and MCC achieved by MLAMP are higher than those achieved by iAMP-2L.

To further demonstrate the power of the MLAMP predictor, we compared it with other classical predictors on an independent dataset \(g^{Ind}\) containing 920 AMPs and 920 non-AMPs. This comparison was used for independent testing in (Thomas et al., 2010; Xiao et al., 2013). The results listed in Table 2 were obtained by MLAMP, iAMP-2L (Xiao et al., 2013) and CAMP (Thomas et al., 2010) on \(g^{Ind}\). As shown in Table 2, the performances achieved by MLAMP is remarkably higher than the performances reported by iAMP-2L (Xiao et al., 2013) and CAMP (Thomas et al., 2010) in all metrics (Sn, Sp, Acc and MCC).
Table 4. Features in the descending order of importance

<table>
<thead>
<tr>
<th>Features in Eq. 13</th>
<th>Description of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_11</td>
<td>Frequency of Methionine</td>
</tr>
<tr>
<td>P_29</td>
<td>Molecular weight (\oplus)</td>
</tr>
<tr>
<td>P_30</td>
<td>Molecular weight (\oplus)</td>
</tr>
<tr>
<td>P_27</td>
<td>PI (\oplus)</td>
</tr>
<tr>
<td>P_28</td>
<td>PI (\oplus)</td>
</tr>
<tr>
<td>P_26</td>
<td>pk2 (\oplus)</td>
</tr>
<tr>
<td>P_21</td>
<td>pk2 (\oplus)</td>
</tr>
<tr>
<td>P_22</td>
<td>Hydrophobicity (\oplus)</td>
</tr>
<tr>
<td>P_17</td>
<td>Frequency of Threonine</td>
</tr>
<tr>
<td>P_20</td>
<td>Frequency of Tyrosine</td>
</tr>
<tr>
<td>P_15</td>
<td>Frequency of Arginine</td>
</tr>
<tr>
<td>P_18</td>
<td>Frequency of Valine</td>
</tr>
<tr>
<td>P_19</td>
<td>Frequency of Tryptophan</td>
</tr>
<tr>
<td>P_14</td>
<td>Frequency of Glutamine</td>
</tr>
<tr>
<td>P_12</td>
<td>Frequency of Asparagine</td>
</tr>
<tr>
<td>P_11</td>
<td>Frequency of Serine</td>
</tr>
<tr>
<td>P_13</td>
<td>Frequency of Proline</td>
</tr>
<tr>
<td>P_24</td>
<td>Frequency of Glutamic Acid</td>
</tr>
<tr>
<td>P_7</td>
<td>Frequency of Histidine</td>
</tr>
<tr>
<td>P_8</td>
<td>Frequency of Lysine</td>
</tr>
<tr>
<td>P_6</td>
<td>Frequency of Isoeucine</td>
</tr>
<tr>
<td>P_6</td>
<td>Frequency of Glycine</td>
</tr>
<tr>
<td>P_5</td>
<td>Frequency of Phenylalanine</td>
</tr>
<tr>
<td>P_3</td>
<td>Frequency of Aspartic Acid</td>
</tr>
<tr>
<td>P_22</td>
<td>Hydrophobicity (\oplus)</td>
</tr>
<tr>
<td>P_20</td>
<td>Frequency of Leucine</td>
</tr>
<tr>
<td>P_10</td>
<td>Frequency of Cysteine</td>
</tr>
<tr>
<td>P_1</td>
<td>Frequency of Alanine</td>
</tr>
</tbody>
</table>

\(\oplus\): parameter of grey model built by molecular weight; \(\oplus\): parameter of grey model built by PI; \(\oplus\): parameter of grey model built by pk2 code; \(\oplus\): parameter of grey model built by hydrophobicity code.

Furthermore, in the second level prediction, MLAMP also obtained better performance than iAMP-2L. Some different metrics were used from single-label classification, in particular-Hamming loss, Accuracy, Precision, Recall and Subset Accuracy (Tsoumakas et al., 2010) were commonly applied in MLC. Table 3 gives the detailed jackknife test results on the AMP dataset \(S^+\) (Eq. 2). Especially MLAMP gained a 0.4846 success rate in the strict assessment of subset accuracy and this performance was 5% higher than that by iAMP-2L.

Why can these metrics be improved so remarkably by using MLAMP? There are two key reasons. The first reason is probably the new peptide feature coding model (see Eq. 13). Table 4 sorts the 30 features in decreasing order after analyzing the benchmark dataset \(S^{\text{benchmark}}\) by the feature selection tool minimal-redundancy-maximal-relevance (mRMR) (Kolde et al., 2016). As shown in Table 4, those features generated by the grey model include more information than amino acids frequency, especially their biochemical properties. And one can draw a conclusion that some physicochemical properties of amino acids may play an important role in AMP, such as molecular weight, PI and PK2. The second reason points to the new ML-SMOTE model. The AMP dataset \(S^+\) is an imbalance MLD, and previous studies did not take it into account. After processing the training dataset \(S^+\) by the ML-SMOTE model, the balance property of the new synthetic training dataset was improved, which can help the machine learning obtained a better performance.

4 Conclusion

Due to increasing antibiotic resistance, AMPs, which are key components of innate immune system, are becoming more and more important in drug development. Efficiently and effectively identifying AMPs and their functional types has become an urgent research topic. The results reported in this study indicate that the novel predictor, MLAMP, provides an accurate and useful tool for researchers to find new infection therapeutics.

MLAMP obtained a better prediction performance than that of a previous method. The primary reason for our good performance is our formulation model’s peptide extraction features. Since AMPs usually have 5–50 amino acids, our model (Eq.13) is good for formulating short peptides. It includes the internal relationship of amino acids sequence in various physical-chemical properties. The second reason is the ML-SMOTE model, which does a good job of handling the lack of balance problem in multi-label datasets. Compared with other methods, the sample synthetized by using ML-SMOTE retains the multi label distributions. It not only accumulates minority samples but also keeps the label density of MLD. In the future, the ML-SMOTE model can be extended to assist with imbalance and multi-label datasets for other problems.

For practical applications, a user-friendly web-server for MLAMP has been established at http://www.jci-bioinfo.cn/MLAMP, which allows users to easily obtain their desired results without the need to follow the complicated mathematical equations involved in developing the predictor. Users can submit a peptide sequence to the webserver and subsequently the webserver will return the predicted result in real time. Alternatively, users can choose the batch prediction by entering their e-mail address and their batch input file of many peptide sequences. They will quickly receive an email showing the predicted results from seconds to hours depending on the number of sequences.

Funding

This work has been supported by the National Natural Science Foundation of China (No. 61462047, 31560316) and by the US National Institutes of Health (GM100701)

Conflict of Interest: none declared.

References


Jia, J. et al. (2016c) Car-PsOpt: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget, 7, 34558-34570.


Liu, A. et al. (2016a) iDHS-EL: identifying Dmase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics, 32, 2411-2418.


