Bioinformatics Advance Access published August 26, 2016

Subject Section

Imbalanced Multi-label Learning for Identifying Antimicrobial Peptides and Their Functional Types

Weizhong Lin[1,2] and Dong Xu[2,*]

[1]nformation Engineering School, Jingdezhen Ceramic Institute, Jingdezhen, 333406, China, [2] Department 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: Dr. John Hancock

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 data sets, 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: A user-friendly web-server for MLAMP was established at http://www.jci-bioinfo.cn/MLAMP.

Contact: 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. found that peptides with antimicrobial and anti-inflammatory activities had therapeutic potential for treatment of acne vulgaris (Popovic, et al., 2012). Yancheva et al. synthesized a novel didepsipeptide with antimicrobial activity against four of five tested bacterial strains of Escherichia coli (Yancheva, et al., 2012). 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 2,544 AMPs from the literature. CAMP (Thomas, et al., 2010; Waghu, et al., 2013) holds 6,756 antimicrobial sequences and 682 3D structures of AMPs, together with prediction and sequence analysis tools. Niarchou et al. (Niarchou, 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. developed LAMP, a database used to aid the discovery and design of AMPs as new antimicrobial agents. The database contains 3,904 natural AMPs and 1,643 synthetic peptides (Zhao, et al., 2013). DBAASP was a manually curated database built by Gogoladze et al., and it collected those peptides for which antimicrobial activities against particular targets have been evaluated experimentally (Gogoladze, et al., 2014).

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 prediction engine (Joseph, et al., 2012; Khosravian, et al., 2013; Lata, et al., 2010; Niarchou, et al., 2013). Besides, nearest neighbor (Wang, et al., 2011) or k-nearest neighbor algorithm (Xiao, et al., 2013), random forests (RFs) (Joseph, et al., 2012), decision tree model (Lira, et al., 2013), and hidden Markov models (HMMs) (Fjell, et al., 2007) were also applied as classifiers. Some predictors were only used to identify whether novel peptides are AMPs (Khosravian, et al., 2013; Lata, et al., 2007; Vishnepolsky and Pirtskhalava, 2014; Wang, et al., 2011). In addition to these simple binary classifiers, there were some multi-class classifiers. Lira et al. (Lira, et al., 2013) created a decision tree model to classify the antimicrobial activities of synthetic peptides into four classes: none, low, medium, and high. Joseph developed ClassAMP to predict the propensity of a peptide sequence to have antibacterial, antifungal, or antiviral activity (Joseph, et al., 2012). Khamis et al., studied 14 AMP families and sub-families. They selected a specific description of AMP amino acid sequence, and identified compositional and physicochemical properties of amino acids to distinguish each AMP family (Khamis, et al., 2015). Furthermore, Xiao et al. proposed a two-level multi-label classifier, iAMP-2L, which identifies not only whether a peptide is an AMP, but also its functional activities (Xiao, et al., 2013).

Although these methods have their own advantages and did play an important role in the research, they have following problems. **Firstly**, most models only identified whether a new sequence is AMP, but not its type. **Secondly**, it is hard to search short peptides in the database because AMPs usually have only 5–50 amino acids. Methods based on Blast search and gene ontology (Lin, et al., 2013) are often ineffective. **Last but not least**, classifying AMPs' functions is a multi-label classification (MLC), especially when the number of AMPs with different activities does not distribute evenly. From APD2 (Wang, et al., 2009), it is seen that antibacterial peptides occupy more than 90% of all AMPs, which is a highly unbalanced MLC. None of aforementioned automatic models considered the unbalanced amounts among various activities.

In the past two decades, the topic of learning from multi-label data sets (MLDs) has drawn significant attention from researchers. Moreover, MLC methods are increasingly applied in various fields, such as semantic annotation of images (Zhang and Zhou, 2007), categorization (Liu and Chen, 2015), and bioinformatics (Cheng, et al., 2014; Chou, 2015; Chou and Shen, 2007; Chou and Shen, 2008; Chou, et al., 2011; Chou, et al., 2012; Sadasivam and Duraiswamy, 2015; Shen and Chou, 2007; Shen and Chou, 2009; Shen and Chou, 2010; Shen and Chou, 2010; Wu, et al., 2012; Yu, et al., 2013). The existing MLC methods can be grouped into two categories: (1) problem transformation methods, which transform MLC either into one or more single-label classification or regression problem, and (2) algorithm adaptation methods, which extend specific learning algorithms in order to handle MLDs directly (Tsoumakas, et al., 2010). Numerous MLC algorithms were proposed, such as Adaboost.MH and Adaboost.MR (Schapire and Singer, 2000), ML-KNN (Zhang and Zhou, 2007), Classifier chains (Read, et al., 2009; Read, et al., 2011), and Multi-label Naïve Bayes (Zhang, et al., 2009).

MLC often has serious issues of unbalanced data sets, in which the numbers of samples from minority classes are substantially fewer than from majority classes. For example, in subcellular localization prediction (Lin, et al., 2013), the number of the cytoplasm proteins is 44 times the number of the melanosome proteins. A similar situation occurs in many studies (Liu and Chen, 2015; Wan, et al., 2012; Wu, et al., 2011; Xiao, et al., 2011). Standard machine learning algorithms often cannot achieve ideal performance when trained on unbalanced data set. One approach to address this issue is to adapt existing classifier learning algorithms to strengthen learning with regard to the minority class(Xu, et al., 2013). Another approach is to artificially sample the class distribution (Dong and Wang, 2011; Luengo, et al., 2011; Tahir, et al., 2012). Combining both approaches can also achieve strong classifiers (Zhang, et al., 2012). Unquestionably, the sampling approach continues to be popular (Chawla, 2010). Various over- and under-sampling methods have been proposed (Bunkhumpornpat, et al., 2009; Chawla, 2010; Chawla, et al., 2002; Chawla, et al., 2003; Dong and Wang, 2011; Gao, et al., 2011; Gao, et al., 2011; Luengo, et al., 2011; Seiffert, et al., 2008; Zhang, et al., 2012). Among them, SMOTE (Synthetic Minority Oversampling TEchinque) (Chawla, 2010; Chawla, et al., 2002) is a state-of-art over-sampling methods. Chawla (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; Jia, et al., 2016; Liu, et al., 2015; Xiao, et al., 2015). In addition, similar approaches have been recently introduced to handle the unbalanced datasets, such as Monte Carlo sampling (Jia, et al., 2016) and fusion ensemble approach (Liu, et al., 2016; Qiu, et al., 2016).

Although the aforementioned methods have some success in addressing unbalanced data sets, 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. (He, et al., 2012) took into account the imbalance in predicting subcellular localization of human proteins. Charte et al. (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 data set $\mathbb{S}^{\mathrm{Bench}}$ used in this study was taken from Xiao et al. (Xiao, et al., 2013). The data set can be formulated as

$$\mathbb{S}^{Bench} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \qquad (1)$$

where $\mathbb{S}^{+}$ contains 879 AMPs, and $\mathbb{S}^{-}$ contains 2,405 non-AMPs. The 879 AMPs are formulated as

$$\mathbb{S}^{+} = \cup_{i=1}^{5} \mathbb{S}_i \tag{2}$$

where $\mathbb{S}_1$ contains 770 antibacterial peptides, $\mathbb{S}_2$ 140 anti-cancer/tumor peptides, $\mathbb{S}_3$ 366 antifungal peptides, $\mathbb{S}_4$ 84 anti-HIV peptides, and $\mathbb{S}_5$ 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

$$P = R_1 R_2 R_3 \ldots R_L \tag{3}$$

where R_i $(1 \le i \le L)$ 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, 2005; Chou, 2001) with the grey model (GM) (Deng, 1989). According to Chou's general PseAAC formula (Chou, 2009; Chou, 2011), the peptides P in Eq. 3 can be represented as

$$\mathbf{P} = [\mathrm{p}_1 \;\; \mathrm{p}_2 \;\; \cdots \;\; \mathrm{p}_\mathrm{k} \;\; \cdots \;\; \mathrm{p}_\Omega]^T \tag{4}$$

where T is a transpose operator, while the subscript Ω is an integer and its value as well as the components p_1, p_2, … 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

$$\mathrm{X}^{(0)} = \left(x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n)\right) \tag{5}$$

is a non-negative original series of real numbers with an irregular distribution. Then

$$\mathrm{X}^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)\right) \tag{6}$$

is viewed as the first-order accumulative generation operation (1-AGO) series for $\mathrm{X}^{(0)}$, and the components in $\mathrm{X}^{(1)}$ are given by

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \quad k = 1,2,\ldots,n \tag{7}$$

The GM(1,1) model can be expressed by the following grey differential equation with one variable:

$$\frac{d\mathrm{X}^{(1)}}{dt} + a\mathrm{X}^{(1)} = b \tag{8}$$

where a and b are elements of parameters vector $\hat{a}$, that is

$$\hat{a} = [a, b]^T \tag{9}$$

In Eq. 8, $-a$ is the developing coefficient and b the influence coefficient. They can be solved using a least square estimator.

$$\hat{a} = [a, b]^T = [B^T B]^{-1} B^T Y \tag{10}$$

where

$$B = \begin{bmatrix} -0.5\left(x^{(1)}(1) + x^{(1)}(2)\right) & 1 \\ -0.5\left(x^{(1)}(2) + x^{(1)}(3)\right) & 1 \\ \vdots & \vdots \\ -0.5\left(x^{(1)}(n-1) + x^{(1)}(n)\right) & 1 \end{bmatrix} \tag{11}$$

$$Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix} \tag{12}$$

The coefficients $-a$ and b should carry some intrinsic information contained in the discrete data sequence $\mathrm{X}^{(0)}$ 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 $(\mathrm{C}^{\alpha} - \mathrm{COOH})$; (3) pk2 (NH3); (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

$$\mathbf{P} = [\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_{20}, \mathrm{p}_{21}, \ldots \mathrm{p}_{30}]^T \tag{13}$$

where p_i $(1 \le i \le 20)$ 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 pk1 numerical values, and so on.

2.3 ML-SMOTE algorithm

In Eq. 2, the AMP function family data set 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

$$Y = \{l_1, l_2, \ldots, l_q\} \tag{14}$$

be a class label set. MLD can be represented as

$$D = \{(x, y) |\; x \in X, y \subseteq Y\} \tag{15}$$

We define the sample set with the j-th $(1 \le j \le q)$ label as

$$D^{(j)} = \left\{\left(x^{(j)}, y^{(j)}\right) |\; (x^{(j)}, y^{(j)}) \in D \;and\; l_j \in y^{(j)}\right\} \tag{16}$$

If $\|D_{j_1}\| \gg \|D_{j_2}\|$, the class l_{j_1} is a majority class and the class l_{j_2} is a minority class.

Different from SMOTE (Chawla, et al., 2002) in a single label data set, the new synthetic instance maybe 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 data set (see Eq. 15) with N samples as

$$D = \left\{t_i = (x_i, y_i) | x_i = \left(x_{i,1}, \dots, x_{i,m}\right), y_i = \ \left(y_{i,1}, \dots y_{i,q}\right), 1 \leq i \leq N \right\} \quad (17)$$

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

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

$$D^{(j)} = \left\{t_i^{(j)} = (x_i^{(j)}, y_i^{(j)}) |\ x_i^{(j)} = \left(x_{i,1}^{(j)}, \dots x_{i,m}^{(j)}\right), y_i^{(j)} = \left(y_{i,1}^{(j)}, \dots y_{i,q}^{(j)}\right),\ and\ y_{i,j}^{(j)} = 1 \right\} \ (1 \leq j \leq q) \qquad (18)$$

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**

```
S = D
MeanSize = $\frac{1}{q}\sum_{j=1}^{q}\|D^{(j)}\|$ (Ti is defined as Eq. 18)
For j = 1→q
 If $\|D^{(j)}\|$ < meanSize
  For each sample $t_i^{(j)}$ in $\|D^{(j)}\|$, do
    Find k-nearest neighbors set knn of sample $t_i^{(j)}$ in $D^{(j)}$
    Randomly select a sample z from knn
   $z = (z_x, z_y)$, where $z_x = (z_{x,1}, \dots z_{x,m}), z_y = (z_{y,1}, \dots, z_{y,q})$
    Get a random vector = $(\underbrace{r_{1,1}, \dots, r_{1,m},}_{r_1}\ \underbrace{r_{2,1}, \dots r_{2,q}}_{r_2})$ , where
   each element of r is a random number between 0 and 1.
    Calculate features of new sample: $v = (1 - r_1).* x_i^{(j)} + r_1.* z_x$.
    Calculate labels of new sample: $u = \mathrm{INT}\left[(1 - r_2).* y_i^{(j)} + r_2.* z_y\right]$
    where .* means array multiplying with element by element,
   and INT[·] means round number.
    Add new sample (v, u) to S
  End for
 End if
End for
```

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

$$u_w = \begin{cases} 1 \quad if\ y_{i,w}^{(j)} = 1\ and\ z_{y,w} = 1 \\ 0 \quad if\ y_{i,w}^{(j)} = 0\ and\ z_{y,w} = 0 \\ 0\ or\ 1, randomly \quad\ if\ y_{i,w}^{(j)} \neq z_{y,w} \end{cases} \quad (1 \leq w \leq q) \qquad (19)$$

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 (Waghu, 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; Wen, et al., 2016) multi-label learning in the Weka (Nicholls, et al., 2016) library And for ECC, the binary and multi-class learners are implemented on the Weka platform using the Random Forest (RF) algorithm (Breiman, 2001).

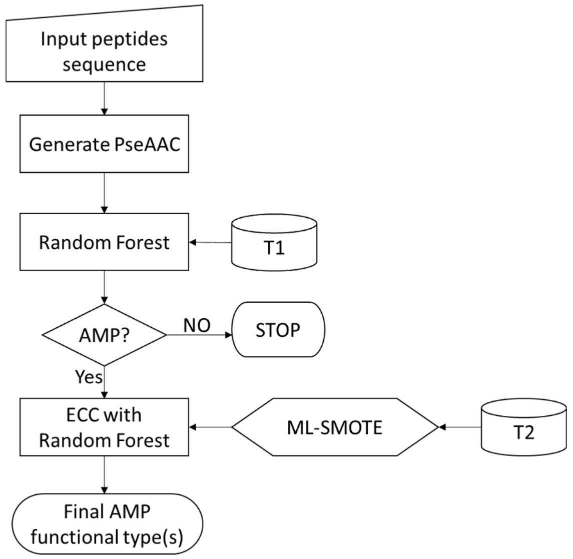

Fig.1. This flowchart shows the training process of MLAMP. T1 represents the data taken from the dataset $\mathbb{S}^{Bench}$ for training the 1st-level predictor; T2 represents those from the dataset $\mathbb{S}^{+}$ for training the 2nd-level predictor.

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: (1) overall accuracy or Acc; (2) Mathew's correlation coefficient or MCC; (3) sensitivity or Sn; and (4) specificity or Sp.

$$\begin{cases} Sn = \frac{TP}{TP+TN} \\ Sp = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+TN+FP+FN} \\ MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases} \qquad (20)$$

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)

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+}, & 0 \leq Sn \leq 1 \\ Sp = 1 - \frac{N_+^-}{N^-}, & 0 \leq Sp \leq 1 \\ Acc = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-}, & 0 \leq Acc \leq 1 \\ MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}}, & 1 \leq MCC \leq 1 \end{cases} \qquad (21)$$

where N$^+$ stands for the total number of AMP samples investigated, whereas N$_-^+$ 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; Jia, et al., 2016; Jia, et al., 2016; Lin, et al., 2014; Liu, et al., 2016; Liu, et al., 2016; Liu, et al., 2016; Qiu, et al., 2016; Xiao, et al., 2016)

If a query peptide is predicted as AMP, the second level of MLAMP will start to classify its functional families. This process belongs to the case of multi-label classification. Hamming loss, Subset Accuracy, Accuracy, Precision and Recall are the mostly used evaluation metrics for the performance of a multi-label classifier (Lin, et al., 2013; Tsoumakas, et al., 2010; Tsoumakas and katakis, 2007; Xiao, et al., 2013). Suppose $\mathbb{L}_k$ is the subset that contains all the labels for the kth sample P_k; $\mathbb{L}_k^*$ 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):

$$\begin{cases} \text{Precision} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\left(\frac{\|\mathbb{L}_\mathrm{k}\cap\mathbb{L}_\mathrm{k}^*\|}{\|\mathbb{L}_\mathrm{K}^*\|}\right) \\ \text{Recall} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\left(\frac{\|\mathbb{L}_\mathrm{k}\cap\mathbb{L}_\mathrm{k}^*\|}{\mathbb{L}_\mathrm{k}}\right) \\ \text{Accuracy} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\left(\frac{\|\mathbb{L}_\mathrm{k}\cap\mathbb{L}_\mathrm{k}^*\|}{\|\mathbb{L}_\mathrm{k}\cup\mathbb{L}_\mathrm{k}^*\|}\right) \\ \text{Subset Accuracy} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\Delta(\mathbb{L}_\mathrm{k},\mathbb{L}_\mathrm{k}^*) \\ \text{Hamming Loss} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}\left(\frac{\|\mathbb{L}_\mathrm{k}\cup\mathbb{L}_\mathrm{k}^*\|-\|\mathbb{L}_\mathrm{k}\cap\mathbb{L}_\mathrm{k}^*\|}{\mathrm{M}}\right) \end{cases} \tag{22}$$

where ‖ ‖ is the operator acting on the set therein to count the number of its elements, and

$$\Delta(\mathbb{L}_k,\mathbb{L}_k^*) = \begin{cases} 1, & \mathit{if\ all\ the\ labels\ in}\ \mathbb{L}_k\ \mathit{are} \\ & \mathit{indentical\ to\ those\ in}\ \mathbb{L}_k^* \\ 0, & \mathit{otherwise} \end{cases} \tag{23}$$

When assessing a predictor, the following three cross-validation methods are often used in the literature: independent data set test, sub-sampling (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 data set. 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:

```
Input: multi-label dataset T={Pi | 1≤ i ≤N}.
Output: predicted label set.
For i: 0→N
    T is divided into testing dataset Ts={Pi},
              and training dataset Tr=T-Ts.
    Generate new training dataset Tr' by using ML-SOMTE on Tr.
    Train model on Tr' by using ECC algorithm.
    Predict the label set of Pi by the model trained above.
End For
```

Table 1 compares the performance of MLAMP with an existing method iAMP-2L in the first-level result on the benchmark $\mathbb{S}^{Bench}$ (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 $\mathbb{S}^{Ind}$ 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 $\mathbb{S}^{Ind}$. 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).

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 $\mathbb{S}^+$ (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.

Tabel 1. Result obtained by MLAMP in identifying AMP in identifying AMP and non-AMP on benchmark $\mathbb{S}^{Bench}$

Predictor	Sn	Sp	Acc	MCC
MLAMP*	77.0%	**94.6%**	**89.9%**	**0.737**
iAMP-2L	**87.1%**	86.0%	86.2%	0.726

* The two parameters, i.e., the number of trees and features used in Random forest were 500 and 6, respectively.

Table 2. Comparison of MLAMP with iAMP-2L and CAMP on the independent data set $\mathbb{S}^{Ind}$

Predictor	Algorithm	Sn	Sp	Acc	MCC
MLAMP	Random forest	**97.3%**	**92.1%**	**94.7%**	**0.895**
iAMP-2L	Fuzzy k-nearest neighbor	97.2%	86.3%	92.2%	0.845
CAMP	Support vector machine	88.4%	66.6%	77.5%	0.55
	Random forest	89.7%	26.0%	57.8%	0.157
	Discriminant analysis	86.6%	64.1%	75.4%	0.508

Table 3. Performance metrics achieved at the 2nd-level by MLAMP on the AMP dataset $\mathbb{S}^+$

Predictor	Hamming loss	Accuracy	Precision	Recall	Subset Accuracy
MLAMP	**0.1595**	**0.6864**	**0.8338**	**0.7631**	**0.4846**
iAMP-2L	0.1640	0.6687	0.8331	0.7570	0.4305

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 data set $\mathbb{S}^{Bench}$ 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 $\mathbb{S}^+$ is an imbalance MLD, and previous studies did not take it into account. After processing the training dataset $\mathbb{S}^+$ by the ML-SMOTE model, the balance property of the new synthetic training data set was improved, which can help the machine learning obtained a better performance.

Table 4. Features in the descending order of importance

Features in Eq. 13	Description of features
P_{11}	Frequency of Methionine
P_{29}	Molecular weight ①
P_{30}	Molecular weight ①
P_{27}	PI ②
P_{28}	PI ②
P_{26}	pk2 ③
P_{25}	pk2 ③
P_{23}	pk1 ④
P_{21}	Hydrophobicity ⑤
P_{17}	Frequency of Threonine
P_{20}	Frequency of Tyrosine
P_{15}	Frequency of Arginine
P_{18}	Frequency of Valine
P_{19}	Frequency of Tryptophan
P_{14}	Frequency of Glutamine
P_{12}	Frequency of Asparagine
P_{16}	Frequency of Serine
P_{13}	Frequency of Proline
P_{4}	Frequency of Glutamic Acid
P_{24}	pk1 ④
P_{7}	Frequency of Histidine
P_{9}	Frequency of Lysine
P_{8}	Frequency of Isoleucine
P_{6}	Frequency of Glycine
P_{5}	Frequency of Phenylalanine
P_{3}	Frequency of Aspartic Acid
P_{22}	Hydrophobicity ⑤
P_{10}	Frequency of Leucine
P_{2}	Frequency of Cysteine
P_{1}	Frequency of Alanine

*①: parameter of grey model built by molecular weight; ②: parameter of grey model built by PI; ③: parameter of grey model built by pk2 code; ④: parameter of grey model built by pk1 code; ⑤: parameter of grey model built by hydrophobicity code.

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 data sets. 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 MLSMOTE model can be extended to assist with imbalance and multi-label data sets 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

Bahar, A.A. and Ren, D. (2013) Antimicrobial Peptides, *Pharmaceuticals*, **6**, 1543-1575.

Breiman, L. (2001) Random forests, *Mach Learn*, **45**, 5-32.

Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. (2009) Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem, *Advances in Knowledge Discovery and Data Mining, Proceedings*, **5476**, 475-482.

Charte, F.*, et al.* (2015) Addressing imbalance in multilabel classification: Measures and random resampling algorithms, *Neurocomputing*, **163**, 3-16.

Chawla, N. (2010) Data Mining for Imbalanced Datasets: An Overview. In Maimon, O. and Rokach, L. (eds), *Data Mining and Knowledge Discovery Handbook*. Springer US, pp. 875-886.

Chawla, N.V.*, et al.* (2002) SMOTE: Synthetic minority over-sampling technique, *J Artif Intell Res*, **16**, 321-357.

Chawla, N.V.*, et al.* (2003) SMOTEBoost: Improving prediction of the minority class in boosting, *Knowledge Discovery in Databases: Pkdd 2003, Proceedings*, **2838**, 107-119.

Chen, W., *et al.* (2016) iACP: a sequence-based tool for identifying anticancer peptides, *Oncotarget*, **7**, 16895-16909.

Cheng, L., *et al.* (2014) Gene Function Prediction Based on the Gene Ontology Hierarchical Structure, *PLoS One*, **9**, e107187.

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

Chou, K.-C. (2009) Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology, *Current Proteomics*, 262-274.

Chou, K.-C. (2013) Some remarks on predicting multi-lable attributes in molecular biosystems, *Mol Biosyst*, 1092-1100.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo amnio acid composition, *Proteins: Structure, Function, and Bioinformatics*, 246-255.

Chou, K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition, *J Theor Biol*, **273**, 236-247.

Chou, K.C. (2015) Impacts of bioinformatics to medicinal chemistry, *Med Chem*, **11**, 218-234.

Chou, K.C. and Shen, H.B. (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, *J Proteome Res*, **6**, 1728-1734.

Chou, K.C. and Shen, H.B. (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, *Nat Protoc*, **3**, 153-162.

Chou, K.C., Wu, Z.C. and Xiao, X. (2011) iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, *PLoS ONE*, **6**, e18258.

Chou, K.C., Wu, Z.C. and Xiao, X. (2012) iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, *Mol Biosyst*, **8**, 629-641.

Chou, K.C. and Zhang, C.T. (1995) Prediction of protein structural classes, *Crit Rev Biochem Mol Biol*, **30**, 275-349.

Conlon, J.M., *et al.* (2014) Potential therapeutic applications of multifunctional host-defense peptides from frog skin as anti-cancer, anti-viral, immunomodulatory, and anti-diabetic agents, *Peptides*, **57**, 67-77.

Deng, J.L. (1989) Introduction to Grey System Theory, *The Journal of Grey System*, 1-24.

Dong, Y.J. and Wang, X.H. (2011) A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets, *Knowledge Science, Engineering and Management*, **7091**, 343-352.

Fjell, C.D., Hancock, R.E. and Cherkasov, A. (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides, *Bioinformatics*, **23**, 1148-1155.

Gao, M., *et al.* (2011) A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, *Neurocomputing*, **74**, 3456-3466.

Gao, M., *et al.* (2011) On Combination of SMOTE and Particle Swarm Optimization Based Radial Basis Function Classifier for Imbalanced Problems, *2011 International Joint Conference on Neural Networks (Ijcnn)*, 1146-1153.

Gaspar, D., Veiga, A.S. and Castanho, M.A. (2013) From antimicrobial to anticancer peptides. A review, *Front Microbiol*, **4**, 294.

Gogoladze, G., *et al.* (2014) DBAASP: database of antimicrobial activity and structure of peptides, *FEMS Microbiol Lett*, **357**, 63-68.

He, J., Gu, H. and Liu, W. (2012) Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites, *PLoS ONE*, **7**, e37155.

Jia, J., *et al.* (2016) iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, *Oncotarget*.

Jia, J., *et al.* (2016) iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets, *Molecules*, **21**, E95.

Jia, J., *et al.* (2016) iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, *Anal Biochem*, **497**, 48-56.

Jia, J., *et al.* (2016) pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, *J Theor Biol*, **394**, 223-230.

Joseph, S., *et al.* (2012) ClassAMP: a prediction tool for classification of antimicrobial peptides, *IEEE/ACM Trans Comput Biol Bioinform*, **9**, 1535-1538.

Khamis, A.M., *et al.* (2015) Distinct profiling of antimicrobial peptide families, *Bioinformatics*, **31**, 849-856.

Khosravian, M., *et al.* (2013) Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods, *Protein Pept Lett*, **20**, 180-186.

Kolde, R., *et al.* (2016) seqlm: an MDL based method for identifying differentially methylated regions in high density methylation array data, *Bioinformatics*.

Lata, S., Mishra, N.K. and Raghava, G.P.S. (2010) AntiBP2: improved version of antibacterial peptide prediction, *Bmc Bioinformatics*, **11 Suppl 1**, S19.

Lata, S., Sharma, B.K. and Raghava, G.P.S. (2007) Analysis and prediction of antibacterial peptides, *Bmc Bioinformatics*, **8**.

Lin, H., *et al.* (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, *Nucleic Acids Res*, **42**, 12961-12972.

Lin, W.Z., *et al.* (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, *Mol Biosyst*, **9**, 634-644.

Lira, F., *et al.* (2013) Prediction of antimicrobial activity of synthetic peptides by a decision tree model, *Appl Environ Microbiol*, **79**, 3156-3159.

Liu, B., *et al.* (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, *Bioinformatics*, **32**, 362-369.

Liu, B., Long, R. and Chou, K.C. (2016) iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, *Bioinformatics*.

Liu, S.M. and Chen, J.H. (2015) A multi-label classification based approach for sentiment classification, *Expert Syst Appl*, **42**, 1083-1093.

Liu, Z., *et al.* (2015) iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition, *Anal Biochem*, **474**, 69-77.

Liu, Z., *et al.* (2016) pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties, *Anal Biochem*, **497**, 60-67.

Luengo, J., *et al.* (2011) Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling, *Soft Comput*, **15**, 1909-1936.

Malmsten, M. (2014) Antimicrobial peptides, *Upsala journal of medical sciences*, **119**, 199-204.

Niarchou, A., *et al.* (2013) C-PAmP: Large Scale Analysis and Database Construction Containing High Scoring Computationally Predicted Antimicrobial Peptides for All the Available Plant Species, *PLoS One*, **8**, e79728.

Nicholls, S.M., Clare, A. and Randall, J.C. (2016) Goldilocks: a tool for identifying genomic regions that are 'just right', *Bioinformatics*.

Popovic, S., *et al.* (2012) Peptides with antimicrobial and anti-inflammatory activities that have therapeutic potential for treatment of acne vulgaris, *Peptides*, **34**, 275-282.

Qiu, W.-R., *et al.* (2016) iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory, *Mol Inform*.

Read, J., *et al.* (2009) Classifier Chains for Multi-label Classification, *Lecture Notes in Computer Science*, **5782**, 254-269.

Read, J., *et al.* (2011) Classifier chains for multi-label classification, *Mach Learn*, **85**, 333-359.

Sadasivam, R. and Duraiswamy, K. (2015) MLDSS: An Algorithm to Mine Multi-Label Disease Spreading Sequence Using Spatio-Time Interval Database, *J Med Imag Health In*, **5**, 17-26.

Schapire, R.E. and Singer, Y. (2000) BoosTexter: A boosting-based system for text categorization, *Mach Learn*, **39**, 135-168.

Seiffert, C., *et al.* (2008) RUSBoost: Improving Classification Performance when Training Data is Skewed, *Int C Patt Recog*, 3650-3653.

Shen, H.B. and Chou, K.C. (2007) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, *Biochem Biophys Res Commun*, **355**, 1006-1011.

Shen, H.B. and Chou, K.C. (2009) Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins, *Protein Pept Lett*, **16**, 1478-1484.

Shen, H.B. and Chou, K.C. (2010) Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins, *J Theor Biol*, **264**, 326-333.

Shen, H.B. and Chou, K.C. (2010) Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites, *J Biomol Struct Dyn*, **28**, 175-186.

Tahir, M.A., Kittler, J. and Yan, F. (2012) Inverse random under sampling for class imbalance problem and its application to multi-label classification, *Pattern Recogn*, **45**, 3738-3750.

Thomas, S., *et al.* (2010) CAMP: a useful resource for research on antimicrobial peptides, *Nucleic Acids Research*, **38**, D774-D780.

Tsoumakas, G., Katakis, I. and Vlahavas, I. (2010) Mining Multi-label Data. In Maimon, O. and Rokach, L. (eds), *Data Mining and Knowledge Discovery Handbook* Springer US, pp. 667-685.

Tsoumakas, G. and katakis, I. (2007) Multi-Label Classification: An Overview, *International Journal Of Data Warehousing and Mining*, **3**, 13.

Vishnepolsky, B. and Pirtskhalava, M. (2014) Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes, *J Chem Inf Model*, **54**, 1512-1523.

Waghu, F.H., *et al.* (2013) CAMP: Collection of sequences and structures of antimicrobial peptides, *Nucleic Acids Res*, **42**, D1154-1158.

Waghu, F.H., *et al.* (2014) CAMP: Collection of sequences and structures of antimicrobial peptides, *Nucleic Acids Res*, **42**, D1154-1158.

Wan, S., Mak, M.W. and Kung, S.Y. (2012) mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines, *BMC Bioinformatics*, **13**, 290.

Wang, G. (2014) Human antimicrobial peptides and proteins, *Pharmaceuticals*, **7**, 545-594.

Wang, G., Li, X. and Wang, Z. (2009) APD2: the updated antimicrobial peptide database and its application in peptide design, *Nucleic Acids Res*, **37**, D933-937.

Wang, P., *et al.* (2011) Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods, *Plos One*, **6**, e18476.

Wen, Q., *et al.* (2016) A gene-signature progression approach to identifying candidate small-molecule cancer therapeutics with connectivity mapping, *BMC Bioinformatics*, **17**, 211.

Wu, Z.C., Xiao, X. and Chou, K.C. (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, *Mol Biosyst*, **7**, 3287-3297.

Wu, Z.C., Xiao, X. and Chou, K.C. (2012) iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins, *Protein Pept Lett*, **19**, 4-14.

Xiao, X., *et al.* (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach, *J Biomol Struct Dyn*, **33**, 2221-2233.

Xiao, X., *et al.* (2013) iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types, *Anal Biochem*, 168-177.

Xiao, X., Wu, Z.C. and Chou, K.C. (2011) iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, *J Theor Biol*, **284**, 42-51.

Xiao, X., *et al.* (2016) iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition, *Oncotarget*.

Xu, Z.Y., *et al.* (2013) Optimization Support Vector Machine, *Front Artif Intel Ap*, **255**, 371-379.

Yancheva, D., *et al.* (2012) Synthesis, structure and antimicrobial activity of 6-(propan-2-yl)-3-methyl-morpholine-2,5-dione, *Journal of Molecular Structure*, **1016**, 147-154.

Yu, G., *et al.* (2013) Protein Function Prediction using Multi-label Ensemble Classification, *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **10**, 1-1.

Zhang, M.-L. and Zhou, Z.-H. (2007) ML-KNN: A lazy learning approach to multi-label learning, *Pattern Recogn*, **40**, 2038-2048.

Zhang, M.L., Pena, J.M. and Robles, V. (2009) Feature selection for multi-label naive Bayes classification, *Inform Sciences*, **179**, 3218-3229.

Zhang, Y., *et al.* (2012) Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions, *Comput Biol Chem*, **36**, 36-41.

Zhao, X., *et al.* (2013) LAMP: A Database Linking Antimicrobial Peptides, *PLoS One*, **8**, e66557.