Predicting Microbe-Drug Association based on Similarity and Semi-Supervised Learning

Lingzhi Zhu, *Jun Wang, Guixiang Li, Xianglong Hu, Bufan Ge and Bohan Zhang

School of Computer and Information Science, Hunan Institute of Technology, Hengyang 421008, China

Abstract: Increasing clinic evidences have showed that microbial communities play important roles in human health and disease. Predicting hidden microbe-drug associations can be helpful in understanding the microbe-drug association mechanisms in clinical treatment, drug discovery, combinations and repositioning. Some computational methods were proposed to predict the associations of microbes and drugs. However, the prediction performance of these methods needs to be improved. In this study, a new computational model (LRLSMDA) is proposed for identifying Microbe-Drug Associations based on the Laplacian Regularized Least Square algorithm. LRLSMDA integrates the chemical structure similarity of drugs and known microbe-drug associations. The microbe Gaussian Interaction Profile (GIP) kernel similarity is computed based on known microbe-drug associations. We compute the drug GIP kernel similarity and the drug chemical structure similarity based on known microbe-drug associations and drug chemical structures. The drug GIP kernel similarity and the drug chemical structure similarity are integrated into a more comprehensive drug similarity matrix by the linear weighted method. Finally, the Laplacian regularized least squares algorithm is applied to predict hidden microbe-drug associations. LRLSMDA has achieved the average Area Under the Curve (AUC) values of 0.8983±0.0019, 0.9043±0.0015 and 0.9095 in 5-fold Cross-Validation (5CV), 10-fold Cross-Validation (10CV) and Leave One Out Cross-Validation (LOOCV), respectively. These experimental results show that the prediction performance of LRLSMDA outperforms three compared models.

Keywords: Microbe-Drug Associations, Similarity, Laplacian Regularized Least Squares, Gaussian Interaction Profile (GIP) Kernel

Introduction

As an important part of the human microbiome, microbes are mainly made up of bacteria, archaea, viruses and fungi etc. Generally speaking, microbes are mainly made up of bacteria, archaea, viruses and fungi etc. Bacteria and viruses are to cause hundreds of human diseases (Geoghegan et al., 2016). Especially for some emerging and epidemic-prone diseases, such as Coronavirus Disease 2019 (COVID-19), Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), directly threaten human health and become the public health concern.

Some researchers think that these diseases can result from the absence of beneficial functions or the introduction of maladaptive functions by invading microbes (Turnbaugh et al., 2007; Methé et al., 2012; Young, 2017). It is also believed that restoring the absence of beneficial functions or eliminating harmful microbial activities is helpful to the treatment of certain diseases (Young, 2017; Huttenhower et al., 2012).

After its discovery in the 1940s, penicillin has been used to restore the absence of beneficial functions or eliminating harmful microbial activities. Millions of people already have been saved by antibiotics from diseases and deaths. Therefore, with the abuse of antibiotics, many bacteria are developing antibiotic resistance, which greatly reduces the efficacy of antibiotics and limits the range of antibiotics. Over 70% of bacteria are resistant to at least one common antibiotic. But at the same time antibiotics are developed rarely and only two antibiotics have been discovered in the past 30 years (Pew Charitable Trusts, 2015). The United Kingdom government predicts that without the...
discovery of new potential antibiotics, 10 million people will die from antibiotic-resistant infections worldwide every year by 2050 (O’neill, 2014). Therefore, drug resistance is a serious threat to public health.

In order to deal with the problem of drug resistance, some scholars proposed two methods to solve this problem: Drug combination and drug repositioning. On the one hand, a medicine application to combat antibiotic resistance is drug combination research (Zimmermann et al., 2007). The first exploration of antimicrobial agents in tuberculosis was performed by using combination drugs (Marshall et al., 1948). Combination drug therapy is being widely used to treat HIV infection and cancer chemotherapy (Vandamme et al., 1998). On the other hand, another method is drug repositioning (Chen et al., 2015), which find novel therapeutic effects of old drugs. For both combinatorial drug treatment and drug repositioning, identifying novel associations between drugs and microbes is their first step (Chen et al., 2016).

Some works show microbes take critical roles in many important biological processes, including an increased toxicity of digoxin (Aarnoudse et al., 2008; Haiser et al., 2014), a reduction of the clearance of morphine and higher morphine AUC inducing virulence in some strains pseudomonas aeruginosa, increasing 221% in simvastatin AUC for homozygote’s (Ong et al., 2012; Voora et al., 2009; Ramsey et al., 2014), altering the activity warfarin (Violi et al., 2016) and an increased toxicity of irnotecan (Guthrie et al., 2017). Identifying associations of microbes and drugs is helpful to throw light on why some respond well to certain drugs, but others suffer severe side-effects. However, to date, only a few microbe-drug associations have been identified (Sun et al., 2018).

In this study, a new model (LRLSMDA) is proposed to identify Microbe-Drug Associations based on the Laplacian Regularized Least Square algorithm. In LRLSMDA, we compute the microbe Gaussian Interaction Profile (GIP) kernel similarity based on known microbe-drug associations to construct the microbe similarity matrix. Then an integrated drug similarity matrix is constructed as follows: First, the chemical structures similarity of drugs is calculated based on the Canonical SMILES of drugs downloaded from Drugbank. Second, we calculate the drug GIP kernel similarity based on known microbe-drug associations. Last, the integrated drug similarity matrix is constructed by the average of the drug GIP kernel similarity and the drug chemical structures similarity. Based on the microbe similarity matrix, the integrated drug similarity matrix and the microbe-drug association matrix, the laplacian regularized least squares algorithm is applied to identify hidden microbe-drug associations.

To confirm the prediction ability of LRLSMDA, we compare LRLSMDA with three compared models. These three models include HGBI (Wang et al., 2013), NBI (Cheng et al., 2012) and SNMF (Wang et al., 2017). Furthermore, we introduce 5-fold cross-validation (5CV), 10-fold Cross-Validation (10CV) and Leave One Out Cross-Validation (LOOCV) to validate whether LRLSMDA is effective in identifying microbe-drug associations. In 5-fold Cross-Validation (5CV), the AUC value of LRLSMDA is 0.8983±0.0019, while the AUC values of HGBI, NBI and SNMF are 0.8516±0.0048, 0.6978±0.0057 and 0.7203±0.0093, respectively. LRLSMDA achieves the better prediction performance than three other models. In 10-fold Cross-Validation (10CV), LRLSMDA is also better as AUC of 0.9030±0.0017, compared with three other models above (HGBI: 0.8721±0.0033, NBI: 0.7098±0.0041 and SNMF: 0.7211±0.0056). In LOOCV, The AUC values of HGBI, NBI, SNMF and LRLSMDA are 0.8873, 0.7199, 0.7622 and 0.9096, respectively. LRLSMDA is also better than three other models.

5CV, 10CV and LOOCV computational experiment results show that LRLSMDA is consistently superior to three other models (HGBI, NBI and SNMF). LRLSMDA is effective in identify hidden miRNA-disease associations.

Materials and Methods

Materials

The dataset of human microbe-drug associations are downloaded from the Microbe-Drug Association Database (MDAD) (Sun et al., 2018). We sort and preprocess these downloaded data and obtain 1152 known microbe-drug associations, 142 microbes and 627 drugs. Let $M = \{m_1, m_2, m_3, \ldots, m_{nm}\}$ denote $nm$ microbes in $M$ and $D = \{d_1, d_2, d_3, \ldots, d_{nd}\}$ represent $nd$ drugs in $D$. Then, $Y$ is $nd \times nm$ columns of the adjacency matrix of microbe-drug associations. If there is a known association between microbe $m_i$ and drug $d_j$, the value of $y_{ij}$ is 1, otherwise is 0. Therefore the benchmark dataset consist of 1,152 known microbe-drug associations and 87,882 unknown microbe-drug associations. This benchmark dataset $\mathcal{Z}$ is represented as follows:

$$\mathcal{Z} = \mathcal{Z}^+ \cup \mathcal{Z}^-,$$

in which $\mathcal{Z}^+$ is 1152 known microbe-drug associations, $\mathcal{Z}^-$ is 87882 unknown microbe-drug associations, respectively.

Methods

Construct the GIP Kernel Similarity Matrix of Microbes

GIP kernel has been successfully applied in many fields (Van Laarhoven et al., 2011; Zhu et al., 2020; Luo et al., 2018). In terms of an assumption that similar microbes
tend to related with similar drugs, the microbe GIP kernel similarity \( KM_{GIP}(m_i,m_j) \) can be computed as:

\[
S_m = KM_{GIP}(m_i,m_j) = \exp\left(-\gamma_m \| y_m - y_m \|^2 \right) \tag{2}
\]

\[
\gamma_m = \gamma'_m \left( \frac{1}{nm} \sum_{i=1}^{m} \| y_m \|^2 \right), \tag{3}
\]

where \( S_m \) is a microbe similarity matrix and \( y_m \) and \( y_m \) are the interaction profiles of microbe \( m_i \) to microbe \( m_j \), respectively. \( \gamma_m \) regulates the normalized kernel bandwidth by the original bandwidth \( \gamma'_m \).

**Construct the Similarity Matrix of Drugs**

For drugs, we compute the drug GIP kernel similarity and the drug chemical structures similarity. According to the microbe GIP kernel similarity calculation method (Zhu et al., 2021), we also compute the drug GIP kernel similarity \( KD_{GIP}(d_i,d_j) \) between drug \( d_i \) and drug \( d_j \) as below:

\[
KD_{GIP}(d_i,d_j) = \exp\left(-\gamma_d \| y_d - y_d \|^2 \right) \tag{4}
\]

\[
\gamma_d = \gamma'_d \left( \frac{1}{md} \sum_{i=1}^{d} \| y_d \|^2 \right), \tag{5}
\]

where, \( y_d \) and \( y_d \) denote the interaction profiles of disease \( d_i \) to disease \( d_j \), respectively. \( \gamma_d \) regulates the normalized kernel bandwidth by the original bandwidth \( \gamma'_d \).

Based on the previous researches, we can use some ways compute the drug similarity. In our study, we introduce the drug chemical structure similarity into LRLSMDA.

The drug chemical structure similarity can be computed by Chemical Development Kit (Steinbeck et al., 2006) based on the chemical structures of drugs in the Canonical Simplified Molecular Input Line Entry Specification (SMILES) (Weininger, 1988). The Canonical Simplified Molecular Input Line Entry Specification format of drugs can be downloaded from Drugbank (Wishart et al., 2018). We compute binary fingerprints of all drugs by Chemical Development Kit. The Tanimoto score (Tanimoto, 1958) of their binary fingerprints is used to measure the chemical structure similarity \( DS_{chem}(d_i,d_j) \).

As shown above, two drug similarity matrices are computed. We combine two drug similarity matrices \( KD_{GIP}(d_i,d_j) \) and \( DS_{chem}(d_i,d_j) \) into a more comprehensive drug similarity matrix \( S_d \) by the linear weighted method:

\[
S_d = \frac{DS_{chem} + KD_{GIP}}{2} \tag{6}
\]

**LRLSMDA for Predicting Microbe-Drug Associations**

The Laplacian Regularized Least Squares (LRLS) algorithm has been successfully applied to identify associations between biological entities. In this study, we present a new model (LRLSMDA) to identify microbe-drug associations via Laplacian Regularized Least Squares algorithm. LRLSMDA is implemented based on the drug chemical structures similarity, the drug GIP kernel similarity and the microbe GIP kernel similarity.

Based on the microbe GIP kernel similarity matrix and the comprehensive drug similarity matrix above, two diagonal matrices \( D_m \) and \( D_d \) can be expressed as follows:

\[
D_m(i,i) = \sum_{j=1}^{n} S_m(i,j) \tag{7}
\]

\[
D_d(j,j) = \sum_{i=1}^{n} S_d(i,j) \tag{8}
\]

Then we normalize these two diagonal matrixes \( D_m \) and \( D_d \) to obtain two normalized laplacian similarity matrixes \( L_m \) and \( L_d \) by the laplacian operation, respectively:

\[
L_m = D_m^{-1/2} (D_m - S_m) D_m^{-1/2} \tag{9}
\]

\[
L_d = D_d^{-1/2} (D_d - S_d) D_d^{-1/2} \tag{10}
\]

Based on this LRLS algorithm, \( FM^* \) and \( FD^* \) are computed using the minimization of the cost functions, respectively:

\[
FM^* = \arg_{FM} \min \left[ \| Y - FM \|_F^2 + \beta_m \cdot tr(FM^T \cdot L_m \cdot FM) \right] \tag{11}
\]

\[
FD^* = \arg_{FD} \min \left[ \| Y - FD \|_F^2 + \beta_d \cdot tr(FD^T \cdot L_d \cdot FD) \right] \tag{12}
\]

in which \( tr(\cdot) \) and \( \| \cdot \|_F \) area matrix trace and the Frobenius norm(Xia et al., 2010), respectively. The trade-off parameters \( \beta_m \) and \( \beta_d \) are set to be 1. Two prediction matrices \( FM^* \) and \( FD^* \) can be computed as

\[
FM^* = S_m \left( S_m + \beta_m \cdot L_m \cdot S_m \right)^{-1} Y \tag{13}
\]

\[
FD^* = S_d \left( S_d + \beta_d \cdot L_d \cdot S_d \right)^{-1} Y^T \tag{14}
\]

Finally, \( FM^* \) and \( FD^* \) are transformed into a prediction matrix with a linear mean method as follows:

\[
F^* = \frac{FM^* + \left( FD^* \right)^T}{2} \tag{15}
\]
Results and Discussion

Performance Evaluation

The prediction performance of LRLSMDA is systematically evaluated by the cross validation framework. In the $k$-fold cross validation, 1152 known microbe-drug associations $S^*$ are divided into $k$ exclusive subsets:

$$S^* = S_1^* \cup S_2^* \cup \ldots \cup S_k^*$$  \hspace{1cm} (16)

With:

$$\emptyset = S_1^* \cap S_2^* \cap \ldots \cap S_k^*$$  \hspace{1cm} (17)

$$|S_1| = |S_2| = \ldots = |S_k|$$  \hspace{1cm} (18)

in which $\cup$ is the symbol of union, $\cap$ is the symbol of intersection and $\emptyset$ is the symbol of the empty set. $S_1^*$ is the first exclusive subset. Each subset (e.g., $S_1^*$) in turn, acts as a test sample and the remaining samples as the training samples. Moreover, all the unknown microbe-drug associations are considered as the candidate associations $k$-fold cross validation is performed 100 times, with the average of predictive results as final results.

In LOOCV, we select each known association as a test sample and the rest known associations as training samples. Moreover, all the unknown microbe-drug associations are selected as the candidate associations.

Each known microbe-drug association is ranked relative to the candidate associations. If the value of this ranking is higher than an assumed threshold, the test sample is correctly predicted.

Comparison with other Models

In order to evaluate the predictive performance of LRLSMDA, we compare it with three other models, namely NBI (Cheng et al., 2012), HGBI (Wang et al., 2013) and SNMF (Wang et al., 2017). NBI is a network-based method to infer new interactions of drugs and targets. HGBI is also a heterogeneous graph inference-based method to infer hidden interactions between drugs and targets. As a matrix factorization-based method, SNMF can predict microbe-drug associations.

5-fold CV, 10-fold CV and LOOCV are used to verify the performance of these models. We can see from Fig. 1 that LRLSMDA is better as AUC of 0.8983±0.0019, compared with three predictive models above (NBI: 0.6978±0.0057, HGBI: 0.8516±0.0048, SNMF: 0.7203±0.0093).

Figure 2 shows that LRLSMDA can obtain the better performance than three other models in 10CV. The AUC value of LRLSMDA is 0.9043±0.0015, while the AUC value of NBI, HGBI and SNMF are 0.7098±0.0041, 0.8721±0.0033 and 0.7211±0.0056, respectively, in 10CV.

In LOOCV, we also compare LRLSMDA with three other models. As shown in Fig. 3, LRLSMDA can achieve the AUC value of 0.9095, while NBI, HGBI and SNMF have 0.7199, 0.8873 and 0.7622 in LOOCV, respectively.
Fig. 2: The AUC curves of four models in 10CV

As we can see from Fig. 1 to 3, LRLSMDA is better than three other models in 5CV, 10CV and LOOCV.

Parameter Analysis

In order to analyze the robustness of LRLSMDA, we quantify the effects of different values of $\gamma_m$ and $\gamma_d$ on the prediction performance of LRLSMDA in 5CV, 10CV and LOOCV, respectively. $\gamma_m$ and $\gamma_d$ regulate the normalized kernel bandwidth by the original bandwidth $\gamma_m'$ and $\gamma_d'$, respectively. Considering that the effect of $\gamma_m'$ is similar to the effect of $\gamma_d'$, $\gamma_m'$ and $\gamma_d'$ are analyzed in our study. We use $\gamma$ denote $\gamma_m'$ and $\gamma_d'$. In Fig. 4, LRLSMDA can obtain AUCs of 0.8821±0.0024, 0.8896±0.0023, 0.8983±0.0019 and 0.8961±0.0024 in 5CV when $\gamma = 2^{-2}$, $2^{-1}$, $2^1$ and $2^2$, respectively. It is
clear that the AUC value of LRLSMDA is $0.8983 \pm 0.0019$ in 5CV when $\gamma$ is equal to $2^1$.

Figure 5 describes an increasing AUC trend of LRLSMDA from $0.886 \pm 0.0014$ to $0.9043 \pm 0.015$ in 10CV, when $\gamma$ increases from $2^{-2}$ to $2^1$. It is obvious for LRLSMDA to make a better performance when the value of parameter is $2^1$.

As shown in Fig. 6, LRLSMDA can obtain AUCs of $0.8892$, $0.8977$, $0.9095$ and $0.9086$ in LOOCV when $\gamma = 2^{-2}$, $2^{-1}$, $2^1$ and $2^2$, respectively. It is obvious for LRLSMDA to make a better performance when $\gamma = 2^1$.

As we can see from Fig. 4 to 6, LRLSMDA makes a better performance in 5CV, 10CV and LOOCV when $\gamma = 2^1$.  

Fig. 4: The AUC curves of LRLSMDA in 5CV when $\sigma$ ranging from $2^{-2}$ to $2^2$.

Fig. 5: The AUC curves of LRLSMDA in 10CV when $\sigma$ ranging from $2^{-2}$ to $2^2$.  

Fig. 6: The AUC curves of LRLSMDA in LOOCV when $\sigma$ ranging from $2^{-2}$ to $2^2$

Conclusion

Increasing evidences have showed that microbes take important roles in human health and disease. Identifying hidden microbe-drug associations is helpful in understanding the microbe-drug association mechanisms in clinical treatment, drug discovery, combinations and repositioning. In our study, LRLSMDA is proposed to predict microbe-drug associations of human. In the model of LRLSMDA, the microbe GIP kernel similarity, the comprehensive drug similarity, the known microbe-drug associations are combined to compute the association score between microbes and drugs. LRLSMDA has achieved the Curve (AUC) values of 0.8983±0.0019, 0.9043±0.0015 and 0.9095 in 5CV, 10CV and LOOCV, respectively, which shows a better performance than three other models.

Acknowledgement

This work is supported in part by the Hengyang Civic Science and Technology Program (2019jh010657, 202010031491), the Scientific Research Foundation of Hunan Provincial Education Department (18B469, 20C0561), the Teaching Reform Research Foundation of Hunan Provincial Education Department (Xiang Jiao Tong [2019] 291-1023), Hunan Provincial Innovation and Entrepreneurship Training Program for Undergraduate (S202011528029, S202011528006), the Scientific Research Foundation of Hunan Institute of Technology(2018HY010).

Author’s Contributions

Lingzhi Zhu: Designed and performed the experiments.  
Jun Wang and Bohan Zhang: Wrote the paper.  
Guixiang Li and Bufan Ge: Participated to collect the materials related to the experiment.  
Jun Wang and Xianglong Hu: Revised the manuscript.

Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

Conflict of Interest

The authors declare that they have no competing interests. The corresponding author affirms that all of the authors have read and approved the manuscript.

References


