frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 03 December 2020
doi: 10.3389/fmolb.2020.603121

Predicting Metabolite-Disease Associations Based on Spy Strategy and ABC Algorithm

Xiujuan Lei, Cheng Zhang and Yueyue Wang*

School of Computer Science, Shaanxi Normal University, Xi'an, China

In recent years, latent metabolite-disease associations have been a significant focus in the biomedical domain. And more and more experimental evidence has been adduced that metabolites correlate with the diagnosis of complex human diseases. Several computational methods have been developed to detect potential metabolite-disease associations. In this article, we propose a novel method based on the spy strategy and an artificial bee colony (ABC) algorithm for metabolite-disease association prediction (SSABCMDA). Due to the fact that there are large parts of missing associations in unconfirmed metabolite-disease pairs, spy strategy is adopted to extract reliable negative samples from unconfirmed pairs. Considering the effects of parameters, the ABC algorithm is utilized to optimize parameters. In relevant cross-validation experiments, our method achieves excellent predictive performance. Moreover, three types of case studies are conducted on three common diseases to demonstrate the validity and utility of SSABCMDA method. Relevant experimental results indicate that our method can predict potential associations between metabolites and diseases effectively.

Keywords: metabolites, disease, associations, spy strategy, ABC algorithm

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Qi Zhao,
University of Science and Technology Liaoning, China
Pu-Feng Du,
Tianjin University, China

****Correspondence:***
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

Received: *05 September 2020*
Accepted: *08 October 2020*
Published: *03 December 2020*

Citation:
Lei X, Zhang C and Wang Y (2020) Predicting Metabolite-Disease Associations Based on Spy Strategy and ABC Algorithm. Front. Mol. Biosci. 7:603121. doi: 10.3389/fmolb.2020.603121

INTRODUCTION

Metabolomics, an important part of systems biology, is a recently and rapidly developed subject following genomics and proteomics, which have entered many fields closely related to human health, such as nutrition and food science, medical development, and, especially, disease diagnosis (Dunn and Ellis, 2005). Accumulating studies have explored the vital roles that metabolites play in the pathogenesis of disease according to changes in the concentration of metabolites. Moreover, the exploration of metabolite–disease associations is meaningful for a deep understanding of the reason a person becomes ill and promotes the diagnosis and treatment of human diseases.

Although many high-throughput metabolomics technologies have been utilized to testify to the metabolite signatures of diseases, which have reached several achievements, such as the Human Metabolome Database (HMDB) (Wishart et al., 2018), unverified metabolite–disease associations are still in the majority. Furthermore, a weakness of experimental determination to identify metabolite-disease associations is that it is extraordinarily laborious and expensive. Accordingly, owing to the high efficiency and reliability of computational approaches (Jiao and Du, 2016; Wu et al., 2020) to identify metabolite–disease associations,

Abbreviations: DAG, Directed Acyclic Graph; GIP, Gaussian interaction profile; LOOCV, Leave-one-out cross validation; TPR, True positive rate; FPR, False positive rate; ROC, Receiver operating characteristic; AUC, Area under the curve.

they have attracted attention from scientific communities in the relevant field. RWRMDA (Hu et al., 2018), the first method for mining the associations between metabolites and diseases, has made progress in developing computational methods in this field. However, the shortcoming of the method is the lack of disease similarity in the construction of the RWRMDA model. The RLS algorithm, whose core framework is regularized least squares, is used in other prediction areas, such as miRNA-disease associations (Chen and Yan, 2014). However, this algorithm uses single similarities which only use biological information as similarity and the performance of it is not stable.

In this article, we put forward a method to predict potential metabolite–disease associations, which utilizes the spy strategy and the artificial bee colony (ABC) algorithm, based on the network consistency projection algorithm (**Figure 1**). First, we select biological properties of diseases and metabolites and integrate them as biological similarity for diseases or metabolites. Simultaneously, the topological properties of diseases and metabolites are also considered when we calculate the final disease similarity. Second, the spy samples from positive samples are utilized to select latent negative samples with suitable thresholds by spy strategy. Third, the optimized parameters are found by utilizing the ABC algorithm. Finally, the network consistency projection algorithm is used to predict the final scores. The area under the curve (AUC) values of the receiver operating characteristic (ROC) are 0.9412 and 0.9355 (average value) in leave-one-out cross validation (LOOCV) and fivefold cross validation, respectively. The case study of tuberculosis, hepatitis, and asthma deeply showed the effectiveness of our method. In summary, the SSABCMDA method could be a useful and effective algorithm for predicting the metabolite–disease associations.

MATERIALS AND METHODS

Metabolite–Disease Associations

The relevant data are extracted from HMDB, DisGeNET (Piñero et al., 2015), and HSDN (Zhou et al., 2014) databases. We firstly extract the disease with DOID and their relevant metabolites in HMDB. Considering integrating the relevant disease similarities, we find the common diseases and their relevant metabolites in DisGeNET and HSDN. Finally, we extract 2,095 experimentally confirmed metabolite–disease pairs, which include 1,401 metabolites and 86 diseases (see **Figure 2**). The unconfirmed metabolite–disease pairs are regarded as unlabeled pairs. In this study, the number of the investigated metabolites and diseases are defined as variables *nm* and *nd*. To distinctly deliver association information, we establish an adjacency matrix A whose size is *nd* rows and *nm* columns. If disease d_i and metabolite m_j are proved to be related, the element *A(i,j)* is set to 1, otherwise 0.

Disease Functional Similarity 1

The scores of disease functional similarity 1 (DFS1) can be calculated under the hypothesis that two diseases which have more similar features are more likely to be linked with similar genes. The associations of diseases and relevant genes are extracted from DisGeNET (Piñero et al., 2015). Subsequently, the Jaccard similarity is used to calculate similarity score between d_i and d_j, which is defined as follows (Gu et al., 2016):

$$DFS1\left(d_i, d_j\right) = \frac{p}{P+q+r} \tag{1}$$

$$G_n\left(d_i\right) = \begin{cases} 1, \text{ if } G_n \text{ is associated with } d_i \text{ and } n \in [1, nd], \\ 0, \text{ otherwise }, \end{cases} \tag{2}$$

where d_i and$d_j$represent two set of diseases. Take d_i as an example, $d_i = [G_1\left(d_i\right), \ldots., G_n\left(d_i\right), ..G_{nd}\left(d_i\right)]$, and p denotes the number of variables with a value of 1 in both $G_n\left(d_i\right)$ and $G_n\left(d_j\right)$ – which means the whole number of genes simultaneously associated with d_i and d_j; q is defined as the number of variables with a value of 1 in $G_n\left(d_i\right)$ and 0 in $G_n\left(d_j\right)$; and r is defined as the number of variables with a value of 0 in $G_n\left(d_i\right)$ and 1 in $G_n\left(d_j\right)$.

Disease Functional Similarity 2

It is assumed that if two diseases obtain a higher score in a symptom-based similarity matrix, they tend to have a more similar function. We extract the relevant symptoms associated with diseases in HSDN. According to previous articles (Zhou et al., 2014; Ma et al., 2016), every disease has its own set that consists of its relevant symptoms, and disease i is taken as an example, which is calculated as follows:

$$D_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,N}) \tag{3}$$

$$w_{i,j} = W_{i,j} \log \frac{nd}{nj} \tag{4}$$

where N is the total number of symptoms, $w_{i,j}$ is defined as the weight of the associations between disease i and symptom j, *nj* denotes the number of diseases that have an association with symptom j, *nd* represents the total number of diseases, $W_{i,j}$ denotes the number of associations of disease i and symptom j,$\log\frac{nd}{nj}$ could balance the weights problem. Then the disease functional similarity (*DFS2*) between the vectors D_i and $D_q$of two diseases i and q is calculated using Equation (5):

$$DFS2\left(d_i, d_q\right) = \cos\left(D_i, D_q\right) = \frac{\sum_{j=1}^{N} D_{i,j} D_{q,j}}{\sqrt{\sum_{j=1}^{N} D_{i,j}^2}\sqrt{\sum_{j=1}^{N} D_{q,j}^2}} \tag{5}$$

Metabolite Function Similarity

This is based on the assumption that two metabolites with functional similarity may have more common relevant enzymes. Using a similar way to obtain DFS2, we calculate the weight vector $M_a$of the metabolites, which is the following:

$$M_a = (w_{a,1}, w_{a,2}, \ldots, w_{a,G}) \tag{6}$$

$$w_{a,b} = W_{a,b} \log \frac{nm}{nb} \tag{7}$$

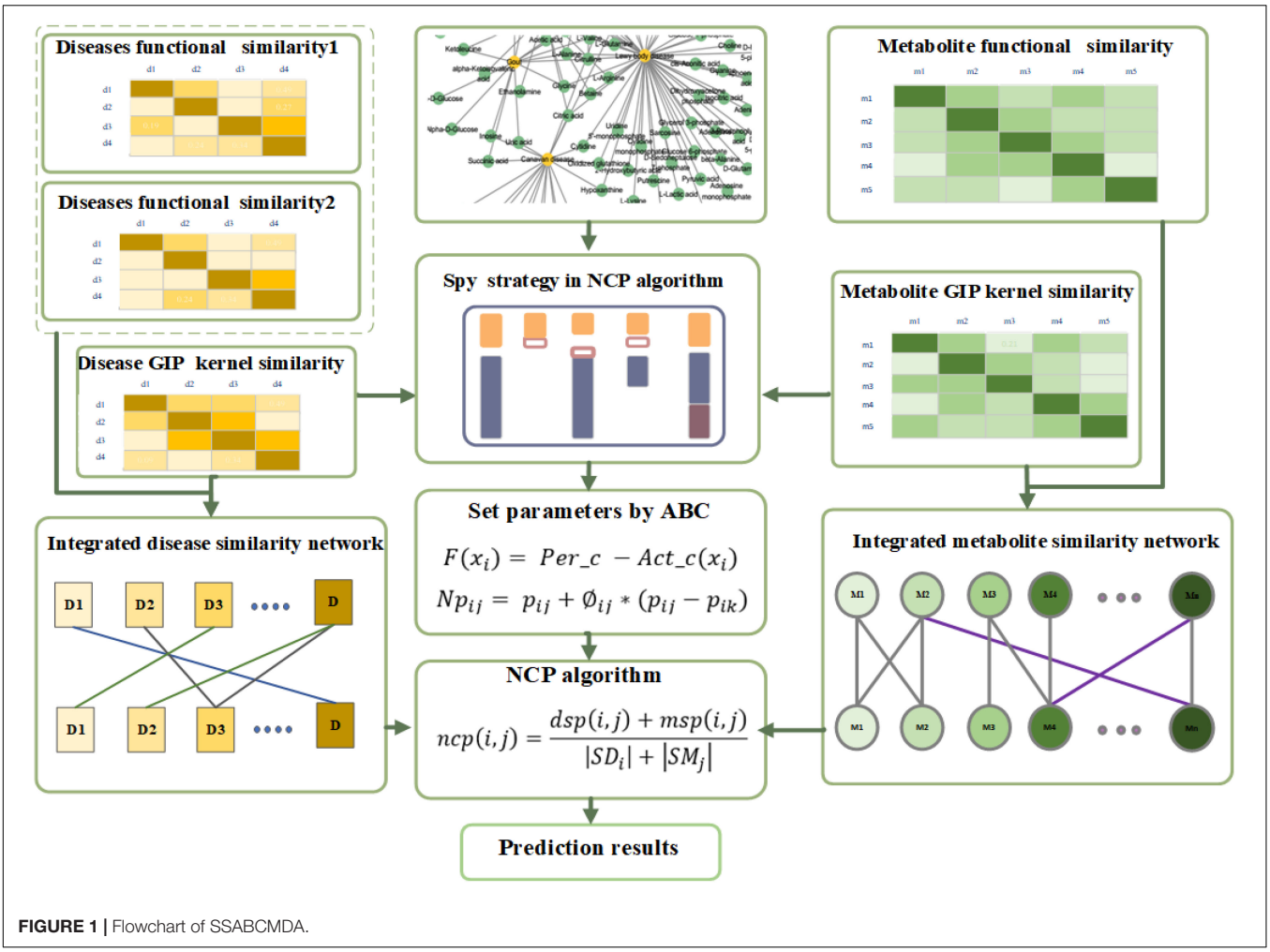

FIGURE 1 | Flowchart of SSABCMDA.

where G is the number of metabolite-related enzymes, $w_{a,b}$ quantifies the strength of the associations between metabolite a and enzyme b, n_b means the number of metabolites associated with enzyme b, nm represents the total number of metabolites, $W_{a,b}$ denotes the number of associations between metabolite a and enzyme b,log $\frac{nm}{nb}$ could balance the weights problem. Finally, the similarity between the vectors M_a and $M_y$of two metabolites a and y is calculated as follows:

$$MFS\left(m_a, m_y\right) = \cos\left(M_a, M_y\right) = \frac{\sum_{b=1}^{G} M_{a,b} M_{y,b}}{\sqrt{\sum_{b=1}^{G} M_{a,b}^2} \sqrt{\sum_{b=1}^{G} M_{y,b}^2}} \quad (8)$$

Gaussian Interaction Profile Kernel Similarity

If we consider the hypothesis that similar metabolites tend to reflect a similar pattern of interaction and non-interaction with diseases, the Gaussian interaction profile (*GIP*) kernel similarity for metabolites and diseases based on the topologic information of known metabolite-disease association network is calculated as follows (Wu et al., 2018; Fang and Lei, 2019):

$$KM\left(m_i, m_j\right) = exp\left(-\omega_m \parallel A\left(:, i\right) - A\left(:, j\right) \parallel^2\right) \quad (9)$$

$$KD\left(d_i, d_j\right) = exp\left(-\omega_d \parallel A\left(i, :\right) - A\left(j, :\right) \parallel^2\right) \quad (10)$$

where $\omega_m$and ω_d denote parameters about kernel bandwidth (Yu et al., 2018), which could be obtained by the normalization operation of the original bandwidth parameter $\omega_m^{'}$, $\omega_d^{'}$ which are set 1, ω_m, ω_d are defined as follows (Jiang et al., 2017):

$$\omega_m = \omega_m^{'} / \left(\frac{1}{nm} \sum_{i=1}^{nm} \parallel A(:, i) \parallel^2\right) \quad (11)$$

$$\omega_d = \omega_d^{'} / \left(\frac{1}{nd} \sum_{i=1}^{nd} \parallel A(i, :) \parallel^2\right) \quad (12)$$

Integrated Similarity for Diseases

In this section, we first integrate two disease functional similarities using the disease biological characteristic similarity

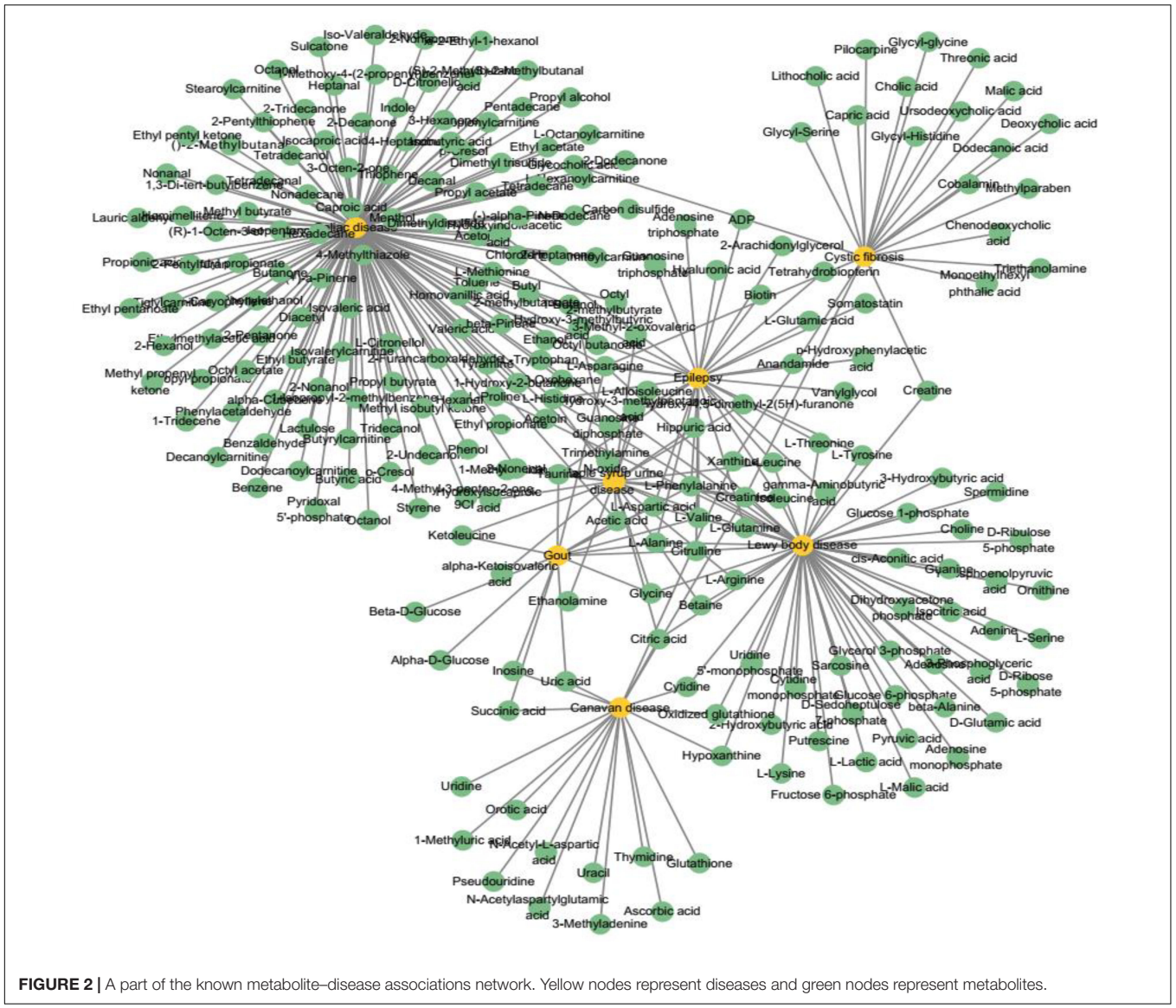

FIGURE 2 | A part of the known metabolite–disease associations network. Yellow nodes represent diseases and green nodes represent metabolites.

(*DB*),which consists of two disease functional similarities, is shown as follows:

$$DB(d_i, d_j) = \begin{cases} DFS1\left(d_i, d_j\right) \text{ if } DFS2\left(d_i, d_j\right) = 0 \\ (1-\alpha)\, DFS2(d_i, d_j) + \alpha DFS1(d_i, d_j) \text{ else} \end{cases} \quad (13)$$

Then the biological and topological characteristics of diseases are integrated, as follows:

$$SD\left(d_i, d_j\right) = \begin{cases} DB\left(d_i, d_j\right) \text{ if } DB\left(d_i, d_j\right) \neq 0 \\ (1-\beta) DB\left(d_i, d_j\right) + \beta KD\left(d_i, d_j\right) \text{ otherwise} \end{cases} \quad (14)$$

Integrated Similarity for Metabolites

The integrated metabolite similarity matrix *SM* consists of metabolite functional similarity and *GIP* kernel similarity. The similarity is defined as below:

$$SM\left(m_i, m_j\right) = \begin{cases} MFS\left(m_i, m_j\right) \text{ if } MFS\left(m_i, m_j\right) \neq 0 \\ (1-\gamma) MFS\left(m_i, m_j\right) + \gamma KM(m_i, m_j) \text{ otherwise} \end{cases} \quad (15)$$

RESULTS

Calculation of Metabolite–Disease Association Prediction Scores

A method named network consistency projection (NCP), which is proposed by Gu et al. (2016) and Bao et al. (2017), is utilized to infer metabolite-disease associations. The main idea for network consistency is that the spatial similarity between metabolite *j* associated metabolites in the metabolite similarity

network and disease i associated metabolites in the metabolite-disease association network or the spatial similarity between disease i associated diseases in the disease similarity network and metabolite j associated diseases in the metabolite–disease association network is positively related to the association between disease i and metabolite j. The potential score between disease i and metabolite j is positively related to the relevant known associations and the spatial similarity in the disease similarity network or metabolite similarity network. There are three steps in the calculation of the predicted score between disease i and metabolite j (Gu et al., 2016; Bao et al., 2017):

First, the scores for metabolite space projection are calculated as follows:

$$msp\left(i,j\right) = \frac{A_i * SM_j}{|A_i|} \tag{16}$$

where $msp\left(i,j\right)$ denotes the projection score of SM_jon A_i, $A_i$represents a vector encoding the associations between disease i and all metabolites in the metabolite–disease association network, SM_j is defined as a vector denoting the similarities between metabolite j, and all metabolites in the metabolite similarity network, $|A_i|$ is the length of vector A_i.

Secondly, the projection scores about diseases should be calculated as follows:

$$dsp\left(i,j\right) = \frac{SD_i * A_j}{|A_j|} \tag{17}$$

where $dsp\left(i,j\right)$ denotes the projection score of DS_i on A_j, A_j represents a vector encoding the associations between metabolite j and all diseases in the metabolite-disease association network, SD_i is defined as a vector denoting the similarities between disease i and all diseases in the disease similarity network, and $\left|A_j\right|$ is the length of vector A_j.

Finally, the predicted scores are integrated relevant scores of the metabolite space projection and disease space projection, which is defined as:

$$ncp\left(i,j\right) = \frac{dsp\left(i,j\right) + msp\left(i,j\right)}{|SD_i| + \left|SM_j\right|} \tag{18}$$

where $ncp\left(i,j\right)$ is the possibility score for disease i and metabolite j, $|SD_i|$denotes the length of DS_i,and $\left|SM_j\right|$represents the length of SM_j.

Spy Strategy

As is generally known, there are many unlabeled metabolite–disease associations in an adjacency matrix, which are regarded as negative training samples most of the time for convenience. But this will cause high false negative rates between predicted associations. Therefore, the spy strategy (Jiang et al., 2017) is utilized to explore the reliable negative samples from the unlabeled metabolite–disease pairs. Spy strategy has several steps. First, 10% spy samples are extracted from the labeled associations, which changes them from 1 to 0. Second, the NCP algorithm and relevant Gaussian kernel similarities are used to get the final score. Then, the score that is the lowest in the spy samples is set to the threshold. If the final score in a candidate sample is lower than the threshold, the relevant value would be set to −1, which is regarded as a reliable negative sample in the association of the metabolite–disease adjacent matrix. Last, the spy samples are repeated 100 times, and the intersection of the reliable negative samples is used as the final reliable negative sample to keep its reliability. The main idea of spy strategy is shown in **Figure 3**.

Parameter Analysis Based on ABC

Testing parameters also play a significant role in prediction performance. Moreover, two articles (Wu et al., 2018; Niu et al., 2020) also point out that a swarm intelligence algorithm can optimize parameters and the ABC algorithm (Karaboga and Akay, 2009) is utilized to get the more suitable parameters α, β, and γ in this article. ABC, which is proposed by Karaboga, is inspired by bee colony behavior. In the ABC search process, the algorithm first needs to be initialized, which includes using the number of positions of the honey sources (*nPo*), the maximum number of iterations (*max_iter*), and the range of parameters. Every honey source position can be regarded as a result (parameter set) $x_i(i = 1, 2, 3, 4,., nPo)$ that is a three-dimensional space ranging from 0 to 1. After initialization, the entire population will repeat the search process with employed, onlooker, and scout bees until the *max_iter* is reached. According to the fitness-function (16), all parameter values are tested, and the best parameter values are found at the end of the algorithm. The fitness function $F(x_i)$ is shown below:

$$F(x_i) = Per\_c - Act\_c(x_i) \tag{19}$$

where Per_c denotes the perfect and ideal forecast result which is set 1, $Act\_c(x_i)$ represents the result aboutx_i, $x_i = \{\alpha, \beta, \gamma$ and $F(x_i)$ represents the honey source cost value. The goal is to obtain a set of suitable parameters whose result could make the $F(x_i)$ turn to be lowest.

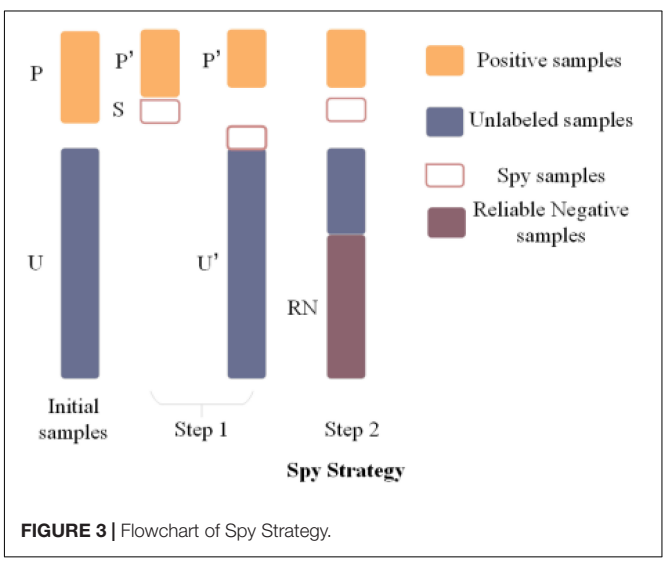

FIGURE 3 | Flowchart of Spy Strategy.

At the beginning of the search process, every employed bee finds a new location of honey source by Equation (19):

$$Np_{ij} = p_{ij} + \emptyset_{ij} * (p_{ij} - p_{ik}) \tag{20}$$

where $k \in [1, npo], j \in [1, D]$ denotes the dimension and $k \neq i$, $\emptyset_{ij} \in [0, 1]$ is random number. As mentioned above, x_i is a set that consists of the values of parameters α, β, and γ. Thus, D is set to 3. After all the employed bees have completed the search, they need to share the relevant information with onlooker bees, and the selection probabilities for each solution are calculated with Equations (20–22):

$$M = \frac{1}{n} * \sum_{i=1}^{n} C_i \tag{21}$$

$$F_i = e^{-\frac{C_i}{M}} i = 1, 2, \ldots, n \tag{22}$$

$$P_i = \frac{F_i}{\sum_{k=1}^{n} F_i} i = 1, 2, \ldots, n \tag{23}$$

where $n \in [1, npo]$ and C_i represents the cost value of the ith honey source, and P_i denotes the selecting probability of the ith honey source. According to probability of every honey source, on-looker bees select honey source and update relevant honey source. When some honey sources are abandoned, the employed bees corresponded to these sources become scout bees. After the convergence criterion was satisfied, we get best cost value of honey source (see **Figure 4**) and the optimal parameters ($\alpha = 0.56, \beta = 0.89, \gamma = 0.6$). In this study, *max_iter*, *nPo*, and the number of employed bees are set to 40, 10, and 10, respectively. In addition, the results of $Act\_c(x_i)$ is calculated by fivefold cross validation (Luo and Xiao, 2017), where we keep the same division of known associations to reduce the impact of other factors on parameter selection.

Performance Evaluation

Leave-one-out cross validation (Liu et al., 2019) and fivefold cross validation (Luo and Xiao, 2017) are used as the evaluation tools for our method. For LOOCV, each association that is confirmed in the database is treated as the test sample while the other known associations are viewed as training samples. In addition, those unconfirmed metabolite–disease pairs are regarded as latent candidate samples. For fivefold cross validation, the known metabolite and disease data are randomly split into five equally sized sets. Each set is retained as the validation samples and the other four sets are treated as the training samples. Similar to the LOOCV, the unconfirmed metabolite–disease pairs are used as the candidate samples. Then, the score for each of the validation samples is ranked against the scores of all the candidate samples. At the same time, we obtain the rank for each association in the test samples. To avoid random error caused by the division of known associations, this procedure is repeated 100 times. According to the results of LOOCV and fivefold cross validation, the AUC – the area under the ROC curve – which is calculated from the true positive rate (TPR) and the false positive rate (FPR), is deemed significant. After LOOCV and fivefold cross validation, SSABCMDA yielded the reliable AUC values of 0.9412 and 0.9355 (average value), respectively, which shows that SSABCMDA presents excellent prediction performance.

The RWRMDA (Hu et al., 2018), RLS algorithm is compared with methods based on the same data in this article. The performance comparison in LOOCV is shown in **Figure 5**, where the AUCs of SSABCMDA, RWRMDA, and the RLS algorithm are 0.9412, 0.6851, 0.7313, respectively. Moreover, SSABCMDA, RWRMDA, and the RLS algorithms gain AUC average values of 0.9355, 0.6738, 0.4371 for fivefold cross validation (see **Figure 6A**). To explore the effects of spy strategy and ABC algorithm, respectively, we compare SSABCMDA; SSABCMDA_1, which doesn't consider spy strategy; and

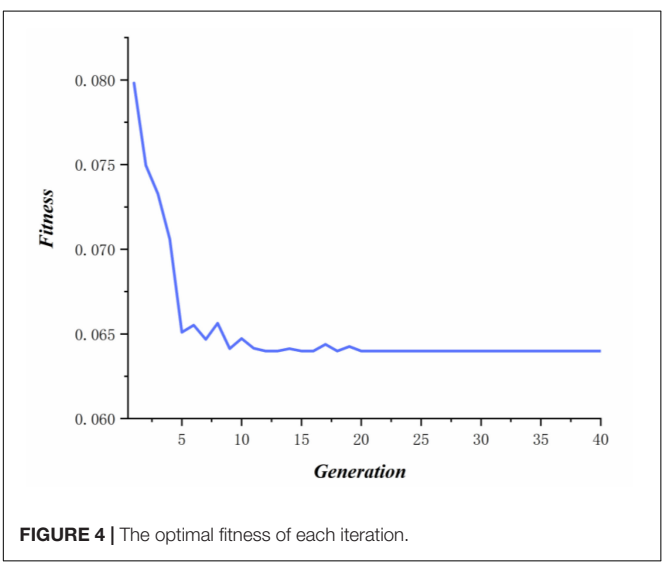

FIGURE 4 | The optimal fitness of each iteration.

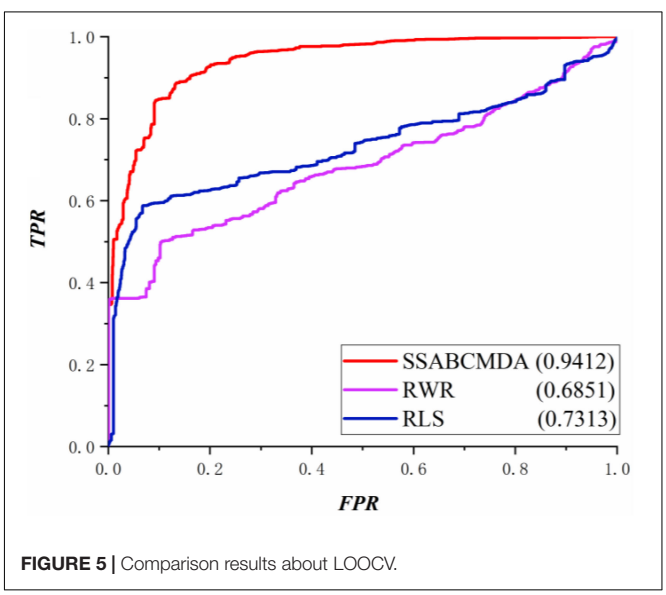

FIGURE 5 | Comparison results about LOOCV.

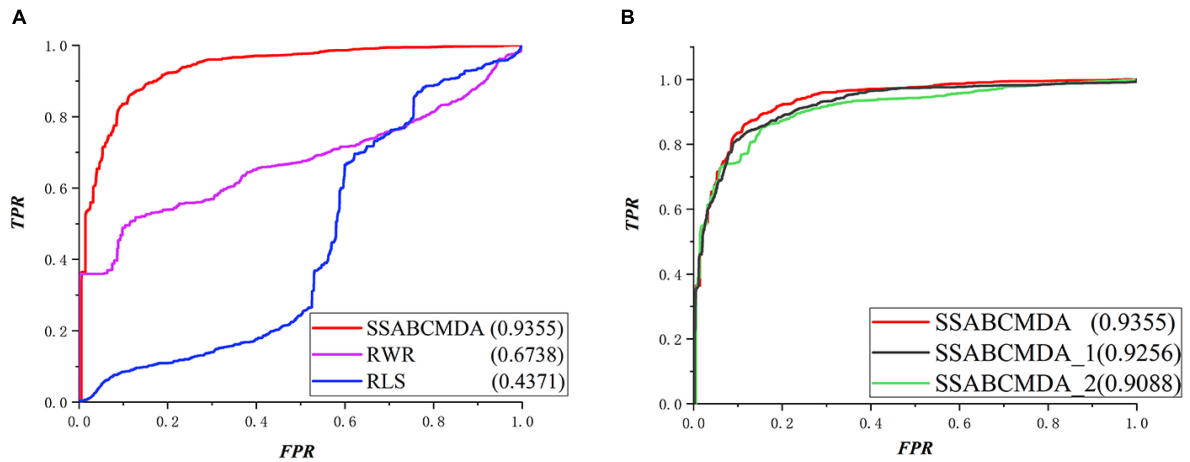

FIGURE 6 | Comparison results about fivefold cross validation. **(A)** Different methods for comparation. **(B)** The combination of different parts in SSABCMDA for comparation.

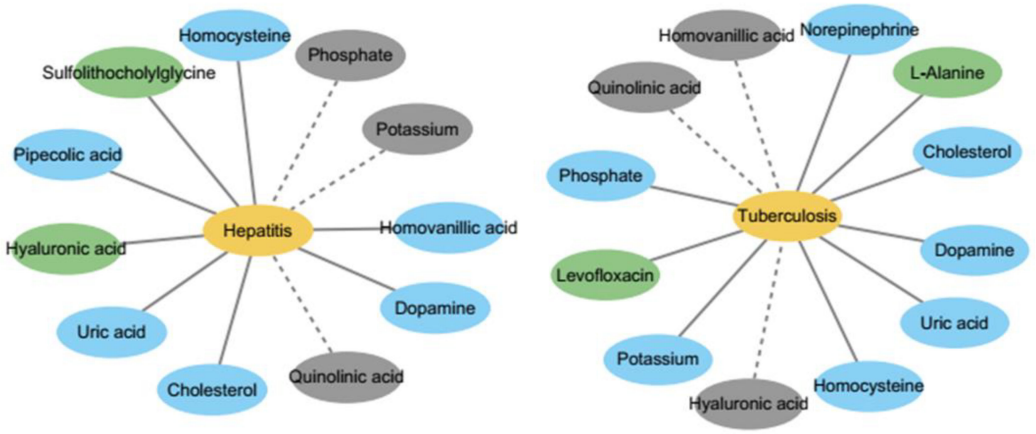

FIGURE 7 | The network of metabolites and diseases. It shows that the top 10 predicted and known metabolites used in this study for two diseases, respectively. The yellow nodes represent diseases and green nodes represent known metabolites which are respective related to two diseases. The blue nodes represent predicted metabolites associated with two disease which are verified by literature, while the gray nodes represent unconfirmed metabolites in top 10 predicted metabolites.

SSABCMDA_2, which only uses random parameters. The relevant results for fivefold cross validation are showed in **Figure 6B**, which indicate that spy strategy and ABC algorithm are effective for predicting performance. As above results showed, we find our method is superior to other methods compared, which indicates that our method is suitable as a reliable biomedical research tool for predicting latent metabolite–disease pairs.

Case Study

In this section, three diseases – tuberculosis, hepatitis, and asthma – are selected for case studies to explore their pathogenic mechanisms with respect to metabolites. Of the top 10 metabolites predicted, 8, 7, and 7 could be verified from the literature for the three diseases. Two diseases and their known and top 10 predicted metabolites are showed in **Figure 7**, which is obvious that the confirmed metabolites in top 10 predicted metabolites can help to study the mechanism of disease from the perspective of metabolism.

Hepatitis is the general name for the liver diseases hepatitis A and B. We conducted a case study of Hepatitis on our calculation method. As shown in **Table 1**, the top 10 metabolites predicted to be interrelated with hepatitis are selected and verified to be correlative. For instance, Uric acid might be useful as a predictive factor for response to therapy for chronic hepatitis (Oh et al., 2017).

Tuberculosis is a chronic infectious disease caused by *Mycobacterium tuberculosis*, which can invade the liver and is most common in pulmonary tuberculosis. There are more than eight million new cases of tuberculosis and 1.3 million deaths (Sharma and Mohan, 2004). We carried out a case study of tuberculosis with our method, and 7 out of top 10 metabolites predicted to be interrelated with tuberculosis are verified to be correlative (see **Table 2**). For instance, the production of NE

TABLE 1 | Candidate metabolites of hepatitis.

Hepatitis		
Rank	**Metabolite name**	**Evidence**
1	Cholesterol	PMID:30600305
2	Uric acid	PMID:28797159
3	Phosphate	----------
4	Dopamine	PMID:30386344
5	Homocysteine	PMID:30063074
6	Quinolinic acid	----------
7	Homovanillic acid	PMID:4817189
8	Potassium	----------
9	Pipecolic acid	PMID:3356409
10	Norepinephrine	PMID:5935605

TABLE 2 | Candidate metabolites of tuberculosis.

Tuberculosis		
Rank	**Metabolite name**	**Evidence**
1	Cholesterol	PMID:29906645
2	Uric acid	PMID:26398460
3	Phosphate	PMID:27105642
4	Dopamine	PMID:25549893
5	Homocysteine	PMID:28936998
6	Quinolinic acid	----------
7	Homovanillic acid	----------
8	Hyaluronic acid	----------
9	Potassium	PMID:30716121
10	Norepinephrine	PMID:27609282

TABLE 3 | Candidate metabolites of asthma.

Asthma		
Rank	**Metabolite name**	**Evidence**
1	Cholesterol	PMID:27839668
2	Uric acid	PMID:26509876
3	Phosphate	PMID:26048149
4	Dopamine	PMID:12055141
5	Homocysteine	----------
6	Quinolinic acid	PMID:23882022
7	Homovanillic acid	PMID:5717841
8	Hyaluronic acid	PMID:24736408
9	Potassium	PMID:11862989
10	Pipecolic acid	----------

(norepinephrine) sharply decreased during advanced infection (Barrios-Payán et al., 2016).

Asthma is a chronic inflammatory disorder arising from heterogenic gene-environment interactions that are still not fully understood (Mims, 2015). A case study of asthma was carried out with our method, and 8 out of top 10 metabolites predicted had associations with asthma (see **Table 3**). For example, hyaluronic acid might be a marker of asthma control, as it correlates with airway resistance and has good sensitivity in the detection of impaired asthma control (Kolesov et al., 1968).

DISCUSSION

In this article, we propose a computational algorithm for metabolite–disease association prediction. To make full use of the information known, we set the known metabolite–disease associations, integrated metabolite similarity, and integrated disease similarity as our input data. The network consistency projection algorithm is utilized as the baseline algorithm. In addition, a spy strategy is utilized to extract negative samples with a high degree of confidence from mixed samples, including potential associations and real negative associations. ABC is introduced to get optimal parameters to improve prediction performance. Moreover, experimental results show reliable evidence that our method is an effective tool to predict metabolite–disease associations. Case studies on three common diseases also give a powerful confirmation to the predictive ability of our method.

The success of our method is due mainly to the following reasons. First, an increasing amount of data known about metabolites and disease has been discovered and confirmed with the development of biological experiments, which are regarded as the basis of predictive data. Second, the network consistency projection as a baseline algorithm guarantees predictive performance. Third, the use of the spy strategy is beneficial to decrease false negative rates of predicted associations. Last, optimal parameters are found quickly with the ABC algorithm, which improves predictive performance.

There are some limitations in the performance of SSABCMDA. At first, although the number of known metabolite–disease associations is larger than before, it is still a small quantity for predictions to obtain sufficiently accurate results. In addition, SSABCMDA depends on the quality of similarity matrices. Some reliable metabolite (disease) similarity matrix from other biological features could be integrated to further expand the algorithm.

DATA AVAILABILITY STATEMENT

These data about metabolite-disease and metabolite-enzyme associations can be found here: https://hmdb.ca/.

AUTHOR CONTRIBUTIONS

XL, CZ, and YW carried out the SSABCMDA method to predict latent associations of metabolites and diseases and participated in its design and drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

Financial support comes from the National Natural Science Foundation of China (61672334, 61972451, and 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

ACKNOWLEDGMENTS

This manuscript is recommended by the 5th Computational Bioinformatics Conference. We thank the conference give our chance. We also thank the financial support which comes from National Natural Science Foundation of China (61672334, 61972451, and 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

REFERENCES

Bao, W., Jiang, Z., and Huang, D. S. (2017). Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinformatics* 18:543. doi: 10.1186/s12859-017-1968-2

Barrios-Payán, J., Revuelta, A., Mata-Espinosa, D., Marquina-Castillo, B., Villanueva, E. B., Gutiérrez, M. E., et al. (2016). The contribution of the sympathetic nervous system to the immunopathology of experimental pulmonary tuberculosis. *J. Neuroimmunol.* 298, 98–105. doi: 10.1016/j.jneuroim.2016.07.012

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4:5501. doi: 10.1038/srep05501

Dunn, W. B., and Ellis, D. I. (2005). Metabolomics: current analytical platforms and methodologies. *Trends Anal. Chem.* 24, 285–294. doi: 10.1016/j.trac.2004.11.021

Fang, Z., and Lei, X. (2019). Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous networ. *Big Data Mining Anal.* 2.4, 248–272. doi: 10.26599/BDMA.2019.9020010

Gu, C., Liao, B., Li, X., and Keqin, L. (2016). Network Consistency Projection for Human miRNA-Disease Associations Inference. *Sci. Rep.* 6:36054. doi: 10.1038/srep36054

Hu, Y., Zhao, T., Zhang, N., Zang, T., Zhang, J., and Cheng, L. (2018). Identifying diseases-related metabolites using random walk. *BMC Bioinformatics* 19:116. doi: 10.1186/s12859-018-2098-1

Jiang, Z. C., Zhen, S., and Bao, W. (2017). "SPYSMDA: SPY strategy-based MiRNA-disease association prediction," in *Proceedings of the International Conference on Intelligent Computing*, eds S. C. Satapathy, J. K. Mandal, V. Bhateja, and M. K. Sanyal (Cham: Springer).doi: 10.1007/978-3-319-63312-1_40

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4:320–330. doi: 10.1007/s40484-016-0081-2

Karaboga, D., and Akay, B. (2009). A comparative study of Artificial Bee Colony algorithm. *Appl. Math. Comput.* 214, 108–132. doi: 10.1016/j.amc.2009.03.090

Kolesov, D., Bol'shakova, T., and Frolov, E. (1968). Excretion of vanillic-amygdalic acid and homovanillic acid in patients with bronchial asthma and pneumonia. *Pediatriia* 47:34.

Liu, Y., Feng, X., Zhao, H., Xuan, Z., and Wang, L. (2019). A novel network-based computational model for prediction of potential LncRNA–disease association. *Int. J. Mol. Sci.* 20:1549. doi: 10.3390/ijms20071549

Luo, J., and Xiao, Q. (2017). A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J. Biomed. Inform.* 66, 194–203. doi: 10.1016/j.jbi.2017.01.008

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2016). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Mims, J. W. (2015). "Asthma: definitions and pathophysiology," in *International Forum of Allergy & Rhinology*, ed. D. W Kennedy (Hoboken, NJ: Wiley Online Library), S2–S6. doi: 10.1002/alr.21609

Niu, M., Zhang, J., Li, Y., Wang, C., Liu, Z., Ding, H., et al. (2020). CirRNAPL: a web server for the identification of circRNA based on extreme learning machine. *Comput. Struct. Biotechnol. J.* 18, 834–842. doi: 10.1016/j.csbj.2020.03.028

Oh, I. S., Won, J. W., Kim, H. J., and Lee, H. W. (2017). Clinical implication of serum uric acid level in pegylated interferon and ribavirin combination therapy for chronic hepatitis C infection. *Korean J. Intern. Med.* 32:1010. doi: 10.3904/kjim.2016.405

Piñero, J., Queraltrosinach, N., Bravo, À, Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* 2015:bav028. doi: 10.1093/database/bav028

Sharma, S., and Mohan, A. (2004). Extrapulmonary tuberculosis. *Indian J. Med. Res.* 120, 316–353.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., et al. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 46, D608–D617. doi: 10.1093/nar/gkx1089

Wu, C., Gao, R., Zhang, D., Han, S., and Zhang, Y. (2018). PRWHMDA: human microbe-disease association prediction by random walk on the heterogeneous network with PSO. *Int. J. Biol. Sci.* 14, 849–857. doi: 10.7150/ijbs.24539

Wu, Q., Wang, Y., Gao, Z., Ni, J., and Zheng, C. (2020). MSCHLMDA: multi-similarity based combinative hypergraph learning for predicting MiRNA-disease association. *Front. Genet.* 11:354. doi: 10.3389/fgene.2020.00354

Yu, S. P., Cheng, L., Qiu, X., Li, G. H., Ding, P. J., and Luo, J. W. (2018). MCLPMDA: a novel method for miRNA-disease association prediction based on matrix completion and label propagation. *J. Cell. Mol. Med.* 23, 1427–1438. doi: 10.1111/jcmm.14048

Zhou, X., Menche, J., Barabási, A. L., and Sharma, A. (2014). Human symptoms–disease network. *Nat. Commun.* 5, 1–10. doi: 10.1038/ncomms5212

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.