A Novel Approach to Predicting Interactions Between Microbes and Hosts

Jie Li (jieli@hit.edu.cn)
Harbin Institute of Technology
Zhuo Chen
Harbin Institute of Technology
Dawei Ma
Harbin Institute of Technology
Li Zhou
Harbin Institute of Technology
Yadong Wang
Harbin Institute of Technology

Research Article

Keywords: human health, KATZHMDA, WBSMDA , NGRHMDA, host interactions, living organisms

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-400365/v1

A novel approach to predicting interactions between microbes and hosts

Jie Li[1,*], Zhuo Chen[1], Dawei Ma[1], Li Zhou[1] and Yadong Wang[1,*]
[1]School of Computer Science and Technology, Harbin Institute of Technology,China
*Correspondence: jieli@hit.edu.cn, ydwang@hit.edu.cn,

Abstract

Increasing evidence shows that microbes are important for the protection of human health and the health of other living organisms. At the same time, microbes can cause other organisms to become sick or even die. Through microbe–host interaction, we can understand intuitively the process and mechanism of host infection by microbes. Several methods are developed to predict microbe–host interactions. However, current methods are limited by the cost of interaction verification experiments and accuracy. Therefore, there is still a need for a rapid and accurate method to predict microbe–host interaction. Here, we proposed a novel method based on Integrated Similarity, KATZ measure, and Within and Between Scores (ISKATZWBS) to predict microbe–host interactions. Experimental results show that the proposed method performs well and the AUCs are 0.946, 0.981, 0.954 on the PHI-base, HPIDB, and HMDAD datasets repectively. Compared with other four state-of the-art methods: KATZHMDA, WBSMDA , NGRHMDA and NCPHMDA, the proposed method has higher prediction accuracy.

Introduction

“Microbe” is a general designation of all microorganisms that are difficult to observe by the naked eye, including bacteria, eukaryotes, archaea, and viruses [1]. On the one hand, the healthy survival of human and other living organisms is inseparable from the role of microbes. Microbes benefit the host with improved metabolic capabilities, protection from pathogens, enhancement of the immune system, and the modulation of gastrointestinal development [2]. Recent work has demonstrated that microbe composition is predictive of the efficacy of cancer immunotherapy [3]. Focusing on microbe–immunity interactions may provide insight into the principles of immune system function while facilitating precision therapeutics for systemic disease [4]. On the other hand, microbes may cause disease in the host or even death. With the development of globalization, global health care is facing an increasing number of severe challenges. For example, on April 21, 2009, the 2009 influenza A(H1N1), caused by influenza virus A, spread rapidly worldwide, and on June 11, 2009, the World Health Organization raised the pandemic alert level to the highest level of 6.2 [5]. As of March 12, 2020, coronavirus disease 2019 (COVID-19) has been confirmed in 125,048

people worldwide, carrying a mortality rate of approximately 3.7%, compared to a mortality rate of less than 1% from influenza [6].

By exploring the interaction between microbes and hosts, we can find effective ways to block transmission and to treat related diseases. With the rapid development of whole genome sequencing, scientists have accumulated a large amount of genetic data. There is a new understanding of biology and biological activities at the molecular level. Scientists use this accumulated data to verify a large number of microbe–host interactions and have stored these interactions in public databases such as PHI-base [7], HPIDB [8], PATRIC [9], PHISTO [10], VirHostNet [11], and Mentha [12]. However, due to the wide variety of proteins and the time and economic cost of verifying microbe–host interactions by experiment, we cannot verify all microbe–host interactions by such methods. Consequently, an increasing number of calculation methods are now used to predict microbe–host interaction relationships with the aim of helping researchers to carry out targeted experiments.

Li et al. [13] used bipartite network projection and introduced an algorithm called the Bipartite Network Module-Based Project (BNMP) to predict pathogen–host association. The method is based on bipartite network modules and integrates the module information of pathogens and hosts respectively into a bipartite network projection model to improve prediction performance. Chen et al. [14] developed a novel computational model of KATZ measure for Human Microbe–Disease Association prediction (KATZHMDA), based on the assumption that functionally similar microbes tend to have similar interaction and non-interaction patterns with non-infectious diseases and vice versa. He et al. [15] developed a novel predictive model of Graph Regularized Non-Negative Matrix Factorization for Human Microbe–Disease Association prediction (GRNMFHMDA). Chen et al. [16] developed the model of Within and Between Scores for MiRNA–Disease Association (WBSMDA prediction) to predict potential miRNAs associated with various complex diseases. WBSMDA can be used to predict unknown miRNA–disease associations.

In this paper, we develop a novel approach to predict associations of microbes and hosts based on Integrated Similarity, KATZ measure, and Within and Between Scores (ISKATZWBS), which is an improvement on KATZHMDA [14] and WBSMDA [16]. In a similar way to KATZHMDA, ISKATZWBS uses Gaussian interaction profile kernel similarity and cosine similarity to obtain the similarity matrix between microorganisms and hosts. The balance factor is then added to the KATZ model, and the results of WBSMDA are used to replace the adjacency matrix of the microbe–host interaction in the KATZ model.

Materials and Method

Known Bacteria–Host Associations

The pathogen–host interactions database PHI-base (http://www.phi-base.org/index.jsp) collects data on bacteria–host interaction by manually reading the literature [7]. Users can download data from PHI-base freely and our known bacteria–host association data is downloaded from PHI-base. There are 267 bacteria, 203 hosts, and 13,802 bacteria–host protein interaction data in the dataset. After data preprocessing, we obtained 819 distinct

confirmed interactions between bacteria and hosts. We constructed the adjacency matrix $A \in R^{n_b \times n_h}$ as the original input data for the experiment, where n_b is the number of bacteria and n_h the number of hosts. $A(i,j) = 1$ indicates that there is a confirmed interaction between the *i*-th bacteria and the *j*-th host; otherwise $A(i,j) = 0$.

Gaussian Interaction Profile Kernel Similarity for Bacteria

Based on the assumption that similar hosts may be infected by the same microbe, we calculate the similarity between bacteria and hosts separately by using Gaussian interaction profile kernel similarity. For the adjacency matrix A, the bacteria–host interaction is expressed as a binary: 1 means there is an interaction between the bacteria and host; otherwise $A = 0$. For a given bacteria b_i, its interaction profile $IP(b_i)$ can be determined by the *i*-th row of the adjacency matrix A. The Gaussian interaction profile kernel similarity between b_i and b_j can be calculated by the following formulae:

$$KB(b_i, b_j) = \exp(-\gamma_b \|IP(b_i) - IP(b_j)\|^2) \quad (1)$$

$$\gamma_b = \gamma_b' / \frac{\sum_{k=1}^{n_b} \|IP(b_k)\|^2}{n_b} \quad (2)$$

In Formula (1), n_b represents the number of bacteria in the dataset and γ_b represents normalized kernel bandwidth based on γ_b' in Formula (2). Following van Laarhoven et al. [17] and Chen et al. [14, 16], we set $\gamma_b' = 1$. From the two formulae above, we know that $KB(b_i, b_i) = 1$ and $0 < KB(b_i, b_j) < 1$. According to Xie et al. [18] and Vanunu et al. [19], when $0 < KB(b_i, b_j) \leq 0.3$, b_i and b_j cannot be considered similar. When $0.6 \leq KB(b_i, b_j) < 1$, b_i and b_j may show significant similarity. Therefore, we use logistic function transformation from $KB(b_i, b_j)$ to $KB^*(b_i, b_j)$ as shown in Formula (3) to make the confidence interval of $KB(b_i, b_j)$ more reasonable.

$$KB^*(b_i, b_j) = 1/(1 + e^{c \times KB(b_i, b_j) + d}) \quad (3)$$

Following Xie et al. [18] and Vanunu et al. [19], we set $c = -15$ and $d = \log(9999)$ so that $KB^*(b_i, b_j) = 0.0001$ when $KB(b_i, b_j) = 0$.

Gaussian Interaction Profile Kernel Similarity for Host

In a similar way to the Gaussian interaction profile kernel similarity of bacteria, for a given host h_i, its interaction profile $IP(h_i)$ can be determined by the *i*-th column of the adjacency matrix A. The Gaussian interaction profile kernel similarity between h_i and h_j can be calculated by the following formulae:

$$KH(h_i, h_j) = \exp(-\gamma_b \|IP(h_i) - IP(h_j)\|^2) \quad (4)$$

$$\gamma_b = \gamma_b' / \frac{\sum_{k=1}^{n_h} \|IP(h_k)\|^2}{n_h} \quad (5)$$

$$KH^*(h_i, h_j) = 1/(1 + e^{c * KH(h_i, h_j) + d}) \quad (6)$$

In the above, n_h is the number of hosts in the dataset. We set $\gamma_b' = 1$, $c = -15$, and $d = \log(9999)$, as before.

Cosine Similarity for Bacteria and Host

The Calculation of cosine similarity (i.e., the calculation of cosine similarity of disease[20], bacteria and host) is based on the assumption that if bacteria b_i and bacteria b_j are similar to each other, then, in the bacteria–host association matrix, pattern $A(:,i)$ (i.e., the i-th column of the adjacency matrix A) and pattern $A(:,j)$ (i.e., the j-th column of adjacency matrix A) should be similar to each other . The same assumption should also be true for host. Therefore, the cosine similarity between bacteria b_i and bacteria b_j can be defined as follows:

$$\mathrm{CB}(b_i, b_j) = \frac{\mathrm{A}(:,i) * \mathrm{A}(:,j)}{||\mathrm{A}(:,i)|| * ||\mathrm{A}(:,j)||} \tag{7}$$

After calculating the cosine similarity between each pair of bacteria–bacteria, the cosine similarity matrix for bacteria $CB(n_b * n_b)$ can be constructed. Similarly, the cosine similarity between host h_i and host h_j is given by:

$$\mathrm{CH}(h_i, h_j) = \frac{\mathrm{A}(i) * \mathrm{A}(j)}{||\mathrm{A}(i)|| * ||\mathrm{A}(j)||} \tag{8}$$

Integrated Similarity for Bacteria and Host

To make full use of the Gaussian interaction profile kernel similarity matrix for bacteria KB and the cosine similarity matrix for bacteria CB, a comprehensive similarity matrix for bacteria $BS(n_b * n_b)$ was constructed by integrating the KB and CB similarity matrices as follows:

$$BS = \frac{KB + CB}{2} \tag{9}$$

In Formula (9), $BS(b_i,b_j)$ represents the integrated similarity between bacteria b_i and bacteria b_j.

In the same way, the Gaussian interaction profile kernel similarity matrix for host KH and the cosine similarity matrix for host CH are integrated into a comprehensive similarity matrix for host $HS(n_h * n_h)$ as follows:

$$HS = \frac{KH + CH}{2} \tag{10}$$

In Formula (10), $HS(h_i,h_j)$ represents the integrated similarity between host h_i and host h_j. As a result of the above, we obtain a comprehensive host similarity matrix HS and a comprehensive bacteria similarity matrix BS, respectively.

WBSMDA

WBSMDA(Within and Between Scores for MiRNA–Disease Association) is an approach to predict Mirna-Disease association by calculating within and between scores [16, 21]. Here we employ it to predict the bacteria-host association. According to WBSMDA, the within-score and between-score of a bacteria–host pair (b_i, h_j) can be defined by the following formulae:

$$S_b^w(b_i, h_j) = \max_{b_k \in b_{h_j}} BS(b_i, b_k) \tag{11}$$

$$S_h^w(b_i, h_j) = \max_{h_k \in h_{b_i}} HS(h_j, h_k) \tag{12}$$

$$S_b^b(b_i, h_j) = \max_{b_k \in b_{\overline{h_j}}} BS(b_i, b_k) \tag{13}$$

$$S_h^b(b_i, h_j) = \max_{h_k \in h_{\overline{b_i}}} HS(h_j, h_k) \tag{14}$$

In the above, b_{h_j} is the group of bacteria that has interaction with h_j in the dataset, h_{b_i} is the group of hosts that has interaction with b_i in the dataset, $b_{\overline{h_j}}$ is the group of bacteria that does not have interaction with h_j in the dataset, and $h_{\overline{b_i}}$ is the group of hosts that does not have interaction with b_i in the dataset. In short, for a given bacteria b_i, within-score is looking for the highest integrated similarity score in the group of bacteria with known interaction with a given host h_j, whereas between-score is looking for the highest integrated similarity score in the group of bacteria with no known interaction with a given host h_j. Similarly, within-score and between-score have the same meanings for a given host h_j.

We then calculate the predicted score $P(b_i, h_j)$ of the interaction between b_i and h_j according to the following formula:

$$P(b_i, h_j) = \frac{S_b^w(b_i, h_j) \times S_h^w(b_i, h_j)}{S_b^b(b_i, h_j) \times S_h^b(b_i, h_j)} \tag{15}$$

In addition, for a given bacteria b that has no known interaction with any host in the dataset, the predicted score $P(b, h_j)$ can be calculated by the following formula:

$$P(b, h_j) = \frac{S_b^w(b, h_j)}{S_b^b(b, h_j)} \tag{16}$$

Similarly, for a given host h that has no known interaction with any bacteria in the dataset, the predicted score $P(b_i, h)$ can be calculated by the following formula:

$$P(b_i, h) = \frac{S_h^w(b_i, h)}{S_h^b(b_i, h)} \tag{17}$$

Finally, we obtain the prediction score matrix $P \in R^{n_b \times n_h}$ by WBSMDA, where $P(i, j)$ is the prediction score for a given bacteria b_i and a given host h_j.

ISKATZWBS

KATZ measure is used for social network prediction [22], disease–gene association prediction [23], lncRNA–disease association prediction [24, 25], and microbe–disease prediction [14, 18]. KATZ is a link prediction method that calculates the similarity of nodes in a heterogeneous

network through random walk. The number of walks between nodes and walk lengths in the network is effective similarity metrics of KATZ Measure. Here, we proposed a new approach IKATZWBS by improving KATZ measure to predict bacteria-host association. The overall process of ISKATZWBS is shown in Figure 1.

First, we calculate the bacteria gaussian interaction profile kernel similarity matrix KB, the bacteria cosine similarity matrix CB, the host gaussian interaction profile kernel similarity matrix KH and the host cosine similarity matrix CH respectively, based on adjacency matrix A. **Then**, in order to make full use of bacteria and host similarity information, two comprehensive similarity matrixes for bacteria and Hosts BS and HS are constructed respectively. **Next**, due to the origin adjacency matrix A was too sparse, a bacteria-host association prediction score matrix P is calculated using WBSMDA which contains more bacteria-host associations. **After that**, we form a heterogeneous network A^* by integrating WBSMDA prediction score matrix P, bacteria integrated similarity matrix BS and host integrated similarity matrix HS. **In addition**, in order to balance the contribution of the integrated similarity for host and the integrated similarity for bacteria during the random walk ,we introduced the balance factor a and b ($a, b \in [0,1]$) to control the contribution of host integrated similarity and bacteria integrated similarity during the process of random walk . **Thus**, the heterogeneous network A^*can be defined as follow:

$$A^* = \begin{bmatrix} a \times BS & P \\ P^T & b \times HS \end{bmatrix} \tag{18}$$

The whole random walking process can be expressed by the formula (19):

$$S = \sum_{l \geq 1} \delta^l A^{*l} = (E - \delta A^*)^{-1} - E \tag{19}$$

In Formula (19), A^{*l} is the l-th power of A^*, the element $A^{*l}(b_i, h_j)$ which is the i-th row and j-th column of A^* means the number of l-length walks between bacteria b_i and host h_j. $\delta \in (0,1)$ is the dampen factor of walks of different length, which means δ^l is the dampen factor of l-length walks. $S \in R^{(n_b+n_h)\times(n_b+n_h)}$ is the final score matrix obtained by the heterogeneous network matrix A^* after l -length walks, and E is the identity matrix. S can be divided into four sub-matrices according to Formula (20):

$$S = \begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} \tag{20}$$

Obviously, $S_1 \in R^{n_b \times n_b}$, $S_2 \in R^{n_b \times n_h}$, $S_3 \in R^{n_h \times n_b}$, and $S_4 \in R^{n_h \times n_h}$. Obviously, S_2 is the final prediction score matrix which we want and the element $S_2(i,j)$ of S_2 is the prediction score of bacteria b_i and host h_j. According to Chen et. [14], when the length of walk l = 2 the KATZ Measure performance is the best. Therefor we also set l = 2 and the whole random walk process can be simplified as follow:

$$S_{l=2} = \delta \times P + \delta^2 (a \times BS \times P + P \times HS \times b) \tag{21}$$

In Formula (21) $S_{l=2} \in R^{n_b \times n_h}$ is the final bacteria – host association prediction result matrix, the element $S_{l=2}(b_i, h_j)$ is the i-th row and j-th column of $S_{l=2}$ which means the final association prediction score between bacteria b_i and host h_j.

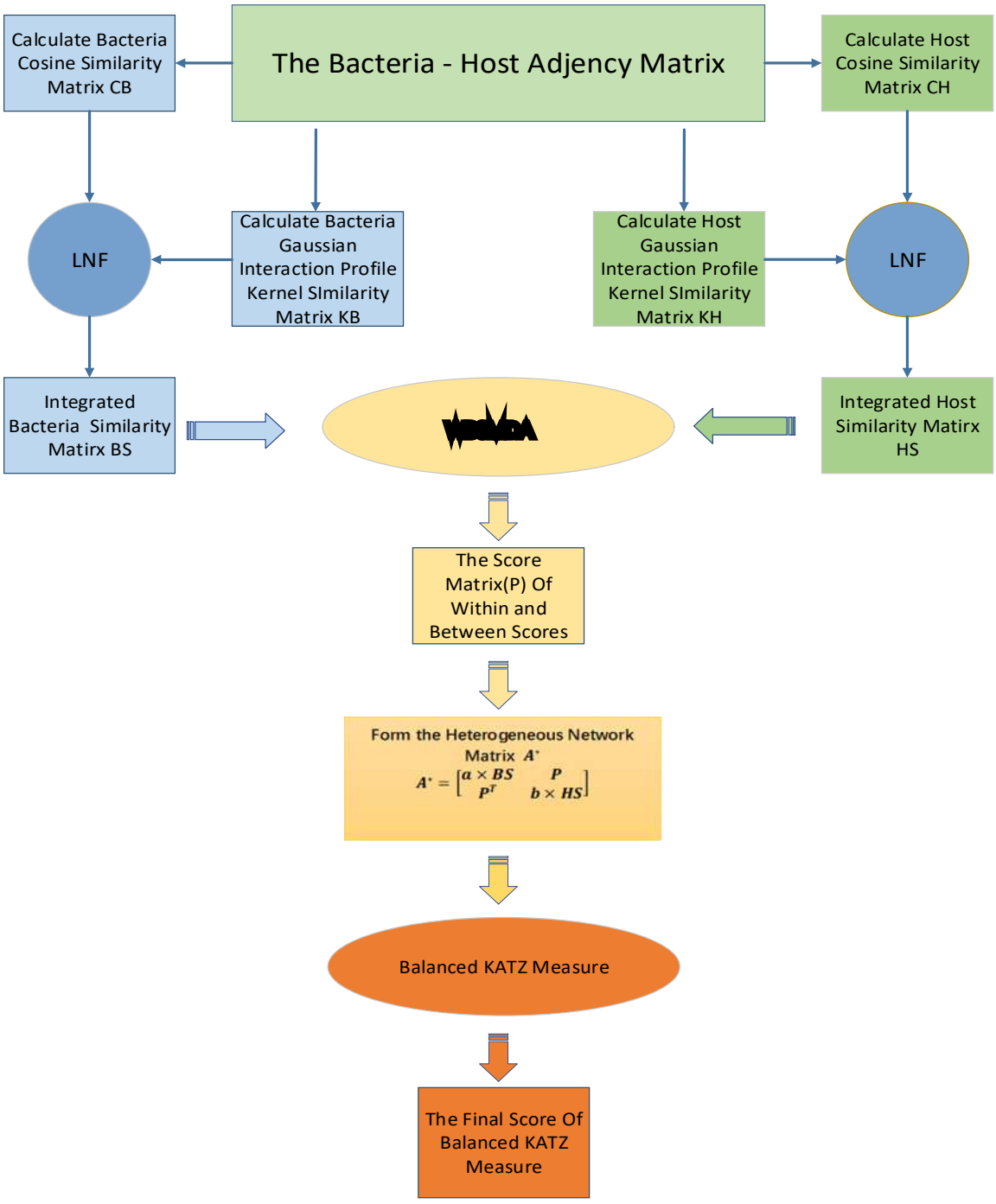

Figure 1. The process of ISKATZWBS

Results

Performance Evaluation

We downloaded bacterial–host interaction data from PHI-base (http://www.phi-base.org/downloadLink.htm) as the verification dataset and used Leave One Out Cross Validation (LOOCV) [26] to test the prediction performance of ISKATZWBS on the verification dataset. In LOOCV, each bacteria–host interaction is used in turn as a test dataset and other bacteria–host interactions are used as a training dataset. In each round of LOOCV, we used ISKATZWBS to make predictions on the training dataset. After LOOCV, we plotted Receiver Operating Characteristic (ROC) curves and Precision Recall (PR) curves to evaluate the performance of the algorithm. In order to display the performance evaluation results more intuitively, we calculated the AUC (the area enclosed by the ROC curve and the coordinate axis) and AUPR (the area enclosed by the PR curve and the coordinate axis), respectively.

With our method, the balance factors a and b are important parameters that affect the final result. We take values a and b from 0 to 1 in steps of 0.1 respectively. We then use data from PHI-base for LOOCV to select the parameters with the best performance results. As shown in Figure 2 and Figure 3, when $a = 0.3$ and $b = 0.1$, the ISKATZWBS prediction has the best performance when AUC=0.947 and AUPR=0.429.

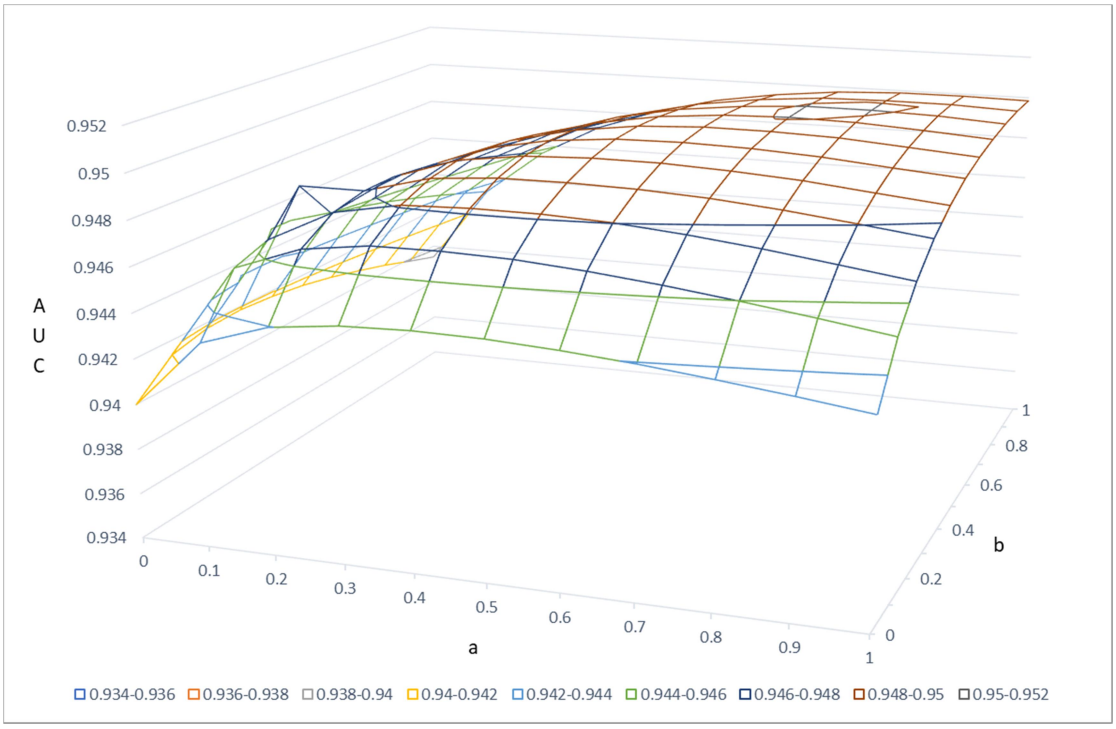

Figure 2. The LOOCV AUC results of ISKATZWBS with different parameters a and b

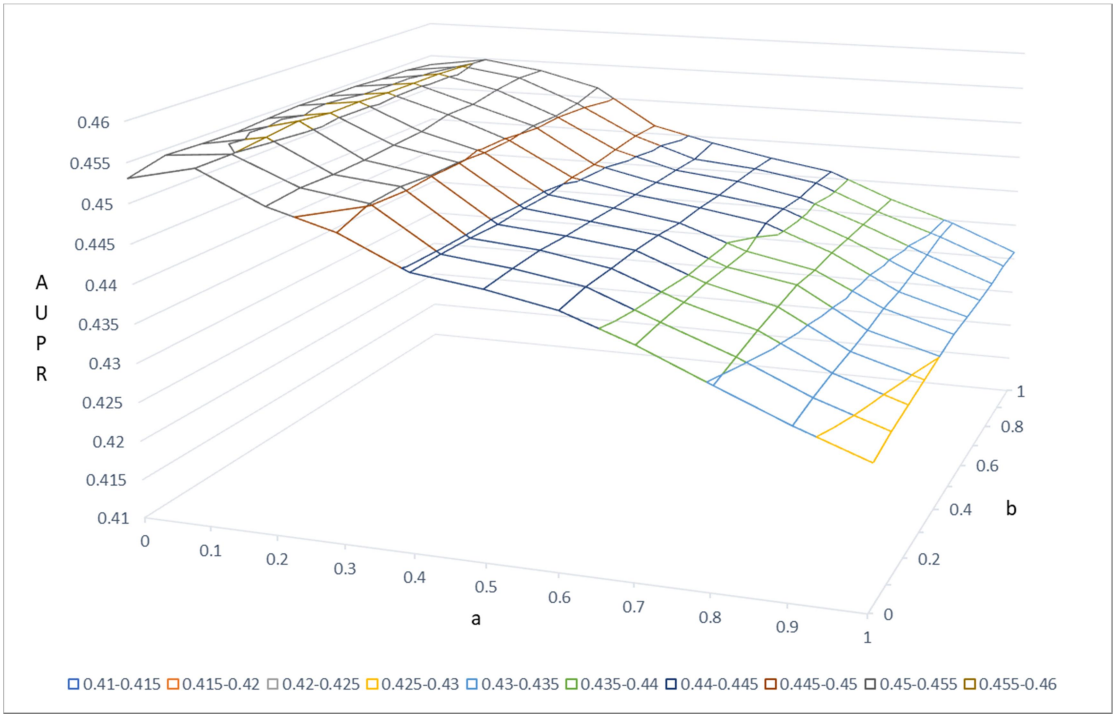

Figure 3. The LOOCV AUPR results of ISKATZWBS with different parameters a and b

Comparison with Existing Methods

We compared ISKATZWBS with four state-of the-art methods – KATZHMDA [14], WBSMDA [16], NGRHMDA[27] and NCPHMDA[29] – by using LOOCV with data from PHI-base [7], HPIDB [8], HMDAD [28]. The data statistics of these databases are shown in Table 1.

Table 1. Data statistics of PHI-base, HPIDB, and HMDAD

Databases	Database content	Data volume (after data preprocessing)
PHI-base	Bacteria–host interaction	270 bacteria, 212 hosts, and 855 distinct bacteria–host interactions.
HPIDB	Virus–host interaction	668 viruses, 66 hosts, and 967 distinct virus–host interactions.
HMDAD	Human microbe–disease association	292 microbes, 39 diseases, and 450 distinct microbe–disease associations.

As above, we used LOOCV to determine the performance of these four methods. The LOOCV AUC results for ISKATZWBS, KATZHMDA, WBSMDA , NGRHMDA and NCPHMDA with data from PHI-base were 0.946, 0.790, 0.833, 0.831 and 0.879 respectively. The LOOCV AUC results for ISKATZWBS, KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA with data from HPIDB were 0.981, 0.943, 0.940, 0.922 and 0.937 respectively. The LOOCV AUC results for ISKATZWBS, KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA with data from HMDAD were 0.954, 0.880, 0.931, 0.862 and 0.929 respectively. The LOOCV AUC results are shown in Figure 4.

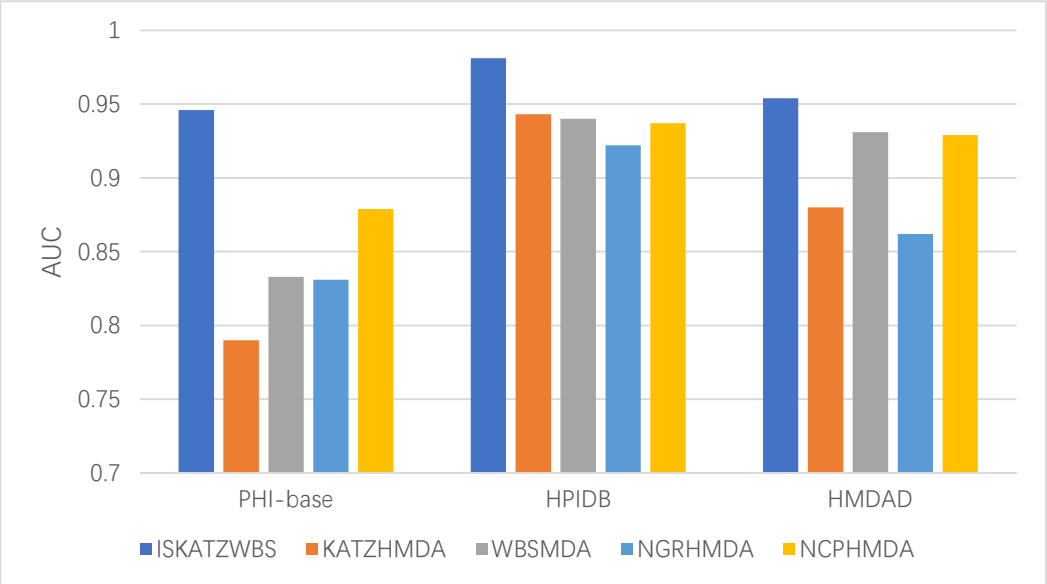

Figure 4. The AUC results of five methods on three data

The LOOCV AUPR results for ISKATZWBS, KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA with data from PHI-base were 0.456, 0.142, 0.271, 0.108 and 0.292 respectively. The LOOCV AUPR results for ISKATZWBS, KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA with data from HPIDB were 0.754, 0.679, 0.635, 0.254 and 0.296 respectively. The LOOCV AUPR results for ISKATZWBS, KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA with data from HMDAD were 0.630, 0.487, 0.483, 0.481 and 0.575 respectively. The LOOCV AUPR results are shown in Figure 5.

Figures 4 and 5 show that ISKATZWBS performs better than other four methods.

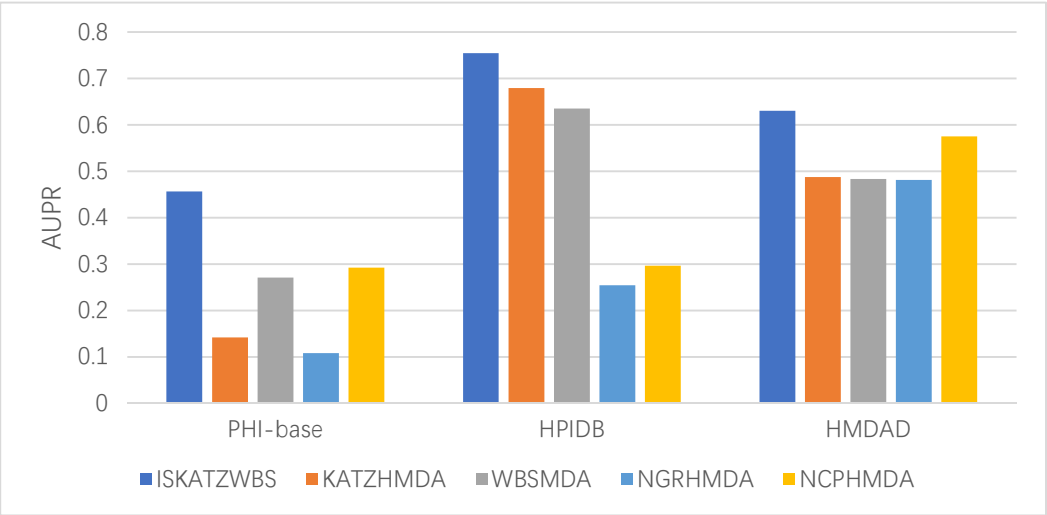

Figure 5. The AUPR results of five methods on three data

.

Validation via Biological Evidence

Due to data mining technology and cost constraints, PHI-base cannot cover all bacterial–host interaction data. In order to further verify the prediction performance of ISKATZWBS, we selected the top 20 potential bacteria–host relationship pairs and searched the literature to find corresponding evidences to prove their potential interaction relationship. We found that 17 of the 20 potential bacteria–host interactions provided biological evidence, as summarized in Table 2 (note that the potential prediction scores ranked at 8, 9, and 13 lacked confirmation in the literature). The details are as follows, with the evidence listed in order of prediction ranking:

1) *Trichophyton mentagrophytes–Mus musculus*. Mackenzie [30] reported an incident in which children were infected with Trichophyton mentagrophytes by mice (Mus musculus). In the subsequent experiment, Trichophyton mentagrophytes were isolated from 104 of 147 mice (Mus musculus).

2) *Vibrio harveyi–Homo sapiens*. Vibrio harveyi is an opportunistic human pathogen that may cause gastroenteritis, severe necrotizing soft-tissue infections, and primary septicemia, with a potentially high rate of lethality. Kang et al. [31] isolated Vibrio harveyi from seawater to study their drug resistance.

3) *Edwardsiella tarda–Homo sapiens*. Jordan et al. [32] found that humans will have symptoms similar to those caused by Salmonellae after being infected with Edwardsiella tarda.

4) *Staphylococcus aureus–Drosophila melanogaster*. Needham et al. [33] found that Drosophila melanogaster can be used to verify the virulence of Staphylococcus aureus.

5) *Bipolaris victoriae–Triticum*. Momtaz et al. [34] infected Triticum through Bipolaris victoriae to discover the relationship between Triticum infected by bipolaris leaf blight.

6) *Venturia inaequalis–Mus musculus*. Following previous reports that the G143A mutation of cytochrome b caused resistance of Venturia inaequalis to a strobilurin-related inhibitor in mice (Mus musculus) mitochondria, Zheng et al. [35] studied the characterization of laboratory mutants of Venturia inaequalis resistant to the strobilurin-related fungicide kresoxim–methyl.

7) *Colletotrichum lagenarium–Nicotiana benthamiana*. Takano et al. [36] found that the Colletotrichum lagenarium wild strain 104-T is able to infect Nicotiana benthamiana, which is not closely related to cucumber.

10) *Verticillium fungicola–Arabidopsis thaliana*. Athey-Pollard et al. [37] compared the percentage homology of Verticillium fungicola and Agaricus bisporus eIF4E amino acid sequences with those isolated from Arabidopsis thaliana and other hosts.

11) *Alternaria citri–Citrus reticulata*. Kono et al. [38] studied host-selective toxins produced by a pathotype of Alternaria citri, a fungus that produces brown spot disease of Dancy tangerine (Citrus reticulata) and other mandarin cultivars.

12) *Escherichia coli–Danio rerio*. Stones et al. [39] described the establishment of a vertebrate model for foodborne EHEC infection, using larval zebrafish (Danio rerio) as a host and the protozoan prey Paramecium caudatum as a vehicle.

14) *Fusarium sporotrichioides–Triticum aestivum*. Asano et al. [40] noted that certain

graminaceous plants such as Zea mays and Triticum aestivum serve as hosts for Fusarium sporotrichioides, and went on to study the molecular interactions between the host plants and F. sporotrichioides.

15) *Beauveria bassiana–Arabidopsis thaliana*. Raad et al. [41] assessed the effects of two Beauveria bassiana strains (BG11 and FRh2) on the growth of Arabidopsis thaliana and its resistance against two herbivore species and a phytopathogen.

16) *Sclerotinia sclerotiorum–Nicotiana tabacum*. Garriz et al. [42] carried out a preliminary evaluation of the potential of polyamine biosynthesis inhibition as a strategy for the control of plant diseases initiated by Sclerotinia sclerotiorum ascospores, using tobacco (Nicotiana tabacum) leaf discs as an experimental system.

17) *Cryptococcus neoformans–Gallus gallus*. Kuroki et al. [43] successfully isolated Cryptococcus neoformans from chicken (Gallus gallus) feces in suburban areas of Thailand.

18) *Acinetobacter nosocomialis–Homo sapiens*. Visca et al. [44] discussed the infection mechanism and threats of Acinetobacter nosocomialis and other Acinetobacter species to humans.

19) *Serratia marcescens–Mus musculus*. Iwaya et al. [45] studied the clinical application and evaluation of rapid and quantitative detection of blood Serratia marcescens by a real-time PCR assay in a mouse (Mus musculus) infection model.

20) *Alternaria alternata–Eriobotrya japonica*. Tziros [46] reported that Alternaria alternata can cause leaf spot and fruit rot on loquat (Eriobotrya japonica) in Greece.

Table 2. The potential bacteria–host interaction prediction results of ISKATZWBS

Order	Bacteria	Hosts	Evidence
1	Trichophyton mentagrophytes	Mus musculus	DOI:10.1080/00362176285190351
2	Vibrio harveyi	Homo sapiens	DOI:10.1016/j.marpolbul.2014.07.008
3	Edwardsiella tarda	Homo sapiens	DOI:10.7326/0003-4819-70-2-283
4	Staphylococcus aureus	Drosophila melanogaster	DOI:10.1099/mic.0.27116-0
5	Bipolaris victoriae	Triticum	DOI:10.3329/jbas.v43i1.42228
6	Venturia inaequalis	Mus musculus	DOI:10.1007/s002940000147
7	Colletotrichum lagenarium	Nicotiana benthamiana	DOI:10.1111/j.1365-2958.2006.05080.x
8	Pyrenopeziza brassicae	Solanum lycopersicum	Unconfirmed
9	Kingella kingae	Caenorhabditis elegans	Unconfirmed
10	Verticillium fungicola	Arabidopsis thaliana	DOI:10.1023/A:1021318524857
11	Alternaria citri	Citrus reticulata	DOI:10.1271/bbb1961.50.1597
12	Escherichia coli	Danio rerio	DOI:10.1128/mSphereDirect.00365-17
13	Streptococcus parauberis	Triticum aestivum	Unconfirmed
14	Fusarium	Triticum aestivum	DOI:10.1186/1477-5956-10-61

	sporotrichioides		
15	Beauveria bassiana	Arabidopsis thaliana	DOI:10.3389/fmicb.2019.00615
16	Sclerotinia sclerotiorum	Nicotiana tabacum	DOI:10.1046/j.1469-8137.2003.00983.x
17	Cryptococcus neoformans	Gallus gallus	DOI:10.1002/yea.1112
18	Acinetobacter nosocomialis	Homo sapiens	DOI:10.1002/iub.600
19	Serratia marcescens	Mus musculus	DOI:10.1016/j.femsle.2005.05.041
20	Alternaria alternata	Eriobotrya japonica	DOI:10.1007/s13314-013-0112-z

Discussion

In this paper, we introduced ISKATZWBS, a novel approach to predicting associations between microbes and hosts, based on integrated similarity, balanced KATZ measure, and within- and between-scores. ISKATZWBS calculates the Gaussian interaction profile kernel similarity and cosine similarity of bacteria and hosts respectively, based on the bacteria–host interaction adjacency matrix. We then integrate the Gaussian interaction profile kernel similarity and cosine similarity of bacteria and hosts, respectively. The within- and between-score matrix obtained by WBSMDA and the bacteria(host) integrated similarity are used in the balanced KATZ measure to predict potential bacteria–host interaction with high accuracy. To prove the universality of our algorithm, we compared ISKATZWBS with KATZHMDA, WBSMDA, NGRHMDA and NCPHMDA by using LOOCV with data from PHI-base, HPIDB and HMDAD. Experimental results show that ISKATZWBS has reliable prediction performance. With minor modifications, ISKATZWBS can also be used to predict other biological associations such as miRNA–disease, lncRNA–disease, drug–target, gene–disease, and drug–cell line interactions.

Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be downloaded from PHI-base (http://www.phi-base.org/downloadLink.htm), HPIDB (https://hpidb.igbb.msstate.edu/about.html), and HMDAD (http://www.cuilab.cn/hmdad).

Declaration

Ethics approval and consent to participate –Not applicable

Consent for publication

Availability of data and materials – Publicly available datasets were analyzed in this study. This data can be downloaded from PHI-base (http://www.phi-base.org/downloadLink.htm), HPIDB

(https://hpidb.igbb.msstate.edu/about.html), and HMDAD (http://www.cuilab.cn/hmdad).

Author Contributions: Conceptualization, Jie Li; methodology, Jie Li, Zhuo Chen and Dawei Ma; software, Zhuo Chen and Dawei Ma; data curation, Zhuo Chen and Dawei Ma; writing—original draft preparation, Jie Li, Zhuo Chen and Dawei Ma; writing—review and editing, Jie Li, Dawei Ma.; supervision, Jie Li; project administration, Jie Li; funding acquisition, Jie Li and Yadong Wang All authors have read and agreed to the published version of the manuscript."

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2016YFC0901905.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be downloaded from PHI-base (http://www.phi-base.org/downloadLink.htm), HPIDB (https://hpidb.igbb.msstate.edu/about.html), and HMDAD (http://www.cuilab.cn/hmdad).

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Acknowledgements –Not applicable

References

[1] WU C, GAO R, ZHANG D, et al. PRWHMDA: human microbe–disease association prediction by random walk on the heterogeneous network with PSO [J]. International Journal of Biological Sciences, 2018, 14(8): 849–57.

[2] VENTURA M, O'FLAHERTY S, et al. Genome-scale analyses of health-promoting bacteria: probiogenomics [J]. Nature Reviews Microbiology, 2009, 7(1): 61–71.

[3] HELMINK B A, KHAN M A W, HERMANN A, et al. The microbiome, cancer, and cancer therapy [J]. Nature Medicine, 2019, 25(3): 377–88.

[4] SPENCER S P, FRAGIADAKIS G K, SONNENBURG J L. Pursuing human-relevant gut microbiota–immune interactions [J]. Immunity, 2019, 51(2): 225–39.

[5] SISTON A M, RASMUSSEN S A, HONEIN M A, et al. Pandemic 2009 influenza A (H1N1) virus illness among pregnant women in the United States [J]. Jama, 2010, 303(15): 1517–25.

[6] MEHTA P, MCAULEY D F, BROWN M, et al. COVID-19: consider cytokine storm syndromes and immunosuppression [J]. The Lancet, 2020, 395(10229): 1033–4.

[7] URBAN M, CUZICK A, SEAGER J, et al. PHI-base: the pathogen–host interactions database [J]. Nucleic Acids Research, 2020, 48(D1): D613–D20.

[8] AMMARI M G, GRESHAM C R, MCCARTHY F M, et al. HPIDB 2.0: a curated database for host–pathogen interactions [J]. Database, 2016.

[9] WATTAM A R, DAVIS J J, ASSAF R, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center [J]. Nucleic Acids Research, 2016, 45(D1): D535–D42.

[10] DURMUŞ TEKIR S, ÇAKıR T, ARDIÇ E, et al. PHISTO: pathogen–host interaction search tool [J]. Bioinformatics, 2013, 29(10): 1357–8.

[11] GUIRIMAND T, DELMOTTE S, NAVRATIL V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data [J]. Nucleic Acids Research, 2014, 43(D1): D583–D7.

[12] CALDERONE A, CASTAGNOLI L, CESARENI G. Mentha: a resource for browsing integrated protein–interaction networks [J]. Nature Methods, 2013, 10(8): 690.
[13] LI J, WANG S, CHEN Z, et al. A bipartite network module-based project to predict pathogen–host association [J]. Frontiers in Genetics, 2020, 10(1357).
[14] CHEN X, HUANG Y A, YOU Z H, et al. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases [J]. Bioinformatics, 2016, btw715.
[15] HE B-S, PENG L-H, LI Z. Human microbe–disease association prediction with graph regularized non-negative matrix factorization [J]. Frontiers in Microbiology, 2018, 9(2560).
[16] CHEN X, YAN C C, ZHANG X, et al. WBSMDA: within and between score for MiRNA–disease association prediction [J]. Scientific Reports, 2016, 6(21106).
[17] TWAN V L, NABUURS S B, ELENA M. Gaussian interaction profile kernels for predicting drug–target interaction [J]. Bioinformatics, 2011, 21.
[18] XIE M, LIU X, LI S. A novel approach based on bipartite network recommendation and KATZ model to predict potential microbe–disease associations [J]. Frontiers in Genetics, 2019, 10(1147).
[19] VANUNU O, MAGGER O, RUPPIN E, et al. Associating genes and protein complexes with disease via network propagation [J]. PLoS Computational Biology, 2010, 6(1).
[20] XIE G, MENG T, LUO Y, LIU Z. SKF-LDA: similarity kernel fusion for predicting lncRNA–disease association. Molecular Therapy – Nucleic Acids, 2019, 18: 45–55.
[21] SHI J Y, LIU Z, HUI Y, et al. Predicting drug–target interactions via within-score and between-score [J]. BioMed Research International, 2015, 2015(1–9).
[22] KATZ L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39–43.
[23] YANG X, GAO L, GUO X, et al. A network based method for analysis of lncRNA–disease associations and prediction of lncRNAs implicated in diseases [J]. PloS One, 2014, 9(1): e87797.
[24] CHEN X. KATZLDA: KATZ measure for the lncRNA–disease association prediction [J]. Scientific Reports, 2015, 5(1): 16840.
[25] CHEN X. Predicting lncRNA–disease associations and constructing lncRNA functional similarity network based on the information of MiRNA [J]. Scientific Reports, 2015, 5(13186).
[26] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]. Proceedings of the IJCAI F, 1995. Montreal, Canada.
[27] Yu-An Huang1, Zhu-Hong You, et al.NGRHMDA: Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model[J]. Translational Medicine, 2017, 15: 209.
[28] MA W, ZHANG L, ZENG P, et al. An analysis of human microbe–disease associations [J]. Briefings in Bioinformatics, 2017, 18(1): 85–97.
[29] Wenzheng Bao, Zhichao Jiang and De-Shuang Huang*:Novel human microbe-disease association prediction using network consistency projection[j]. Briefings in Bioinformatics, 2017, 18(16): 173-181
[30] MACKENZIE D. Trichophyton mentagrophytes in mice: infections of humans and incidence amongst laboratory animals [J]. Sabouraudia, 1962, 1(3): 178–82.
[31] KANG C-H, KIM Y, OH S J, et al. Antibiotic resistance of vibrio harveyi isolated from seawater in Korea [J]. Marine Pollution Bulletin, 2014, 86(1–2): 261–5.
[32] JORDAN G W, HADLEY W K. Human infection with edwardsieila tarda [J]. Annals of Internal Medicine, 1969, 70(2): 283–8.

[33] NEEDHAM A J, KIBART M, CROSSLEY H, et al. Drosophila melanogaster as a model host for Staphylococcus aureus infection [J]. Microbiology, 2004, 150(7): 2347–55.
[34] MOMTAZ M S, SHAMSI S, DEY T K. Association of bipolaris and drechslera species with bipolaris leaf blight (BPLB) infected wheat leaves [J]. Journal of Bangladesh Academy of Sciences, 2019, 43(1): 11–6.
[35] ZHENG D, OLAYA G, KÖLLER W. Characterization of laboratory mutants of venturia inaequalis resistant to the strobilurin-related fungicide kresoxim–methyl [J]. Current Genetics, 2000, 38(3): 148–55.
[36] TAKANO Y, TAKAYANAGI N, HORI H, et al. A gene involved in modifying transfer RNA is required for fungal pathogenicity and stress tolerance of colletotrichum lagenarium [J]. Molecular Microbiology, 2006, 60(1): 81–92.
[37] ATHEY-POLLARD A, KIRBY M, POTTER S, et al. Comparison of partial sequence of the cap binding protein (eIF4E) isolated from agaricus bisporus and its pathogen verticillium fungicola [J]. Mycopathologia, 2003, 156(1): 19–23.
[38] KONO Y, GARDNER J, SUZUKI Y, et al. Studies on host-selective toxins produced by a pathotype of alternaria citri causing brown spot disease of mandarins [J]. Agricultural and Biological Chemistry, 1986, 50(6): 1597–606.
[39] STONES D H, FEHR A G, THOMPSON L, et al. Zebrafish (danio rerio) as a vertebrate model host to study colonization, pathogenesis, and transmission of foodborne escherichia coli O157 [J]. mSphere, 2017, 2(5): e00365-17.
[40] ASANO T, KIMURA M, NISHIUCHI T. The defense response in arabidopsis thaliana against fusarium sporotrichioides [J]. Proteome Science, 2012, 10(1): 61.
[41] RAAD M, GLARE T R, BROCHERO H L, et al. Transcriptional reprogramming of arabidopsis thaliana defence pathways by the entomopathogen beauveria bassiana correlates with resistance against a fungal pathogen but not against insects [J]. Frontiers in Microbiology, 2019, 10(615).
[42] GÁRRIZ A, DALMASSO M C, MARINA M, et al. Polyamine metabolism during the germination of sclerotinia sclerotiorum ascospores and its relation with host infection [J]. New Phytologist, 2004, 161(3): 847–54.
[43] KUROKI M, PHICHAICHUMPON C, YASUOKA A, et al. Environmental isolation of cryptococcus neoformans from an endemic region of HIV-associated cryptococcosis in Thailand [J]. Yeast, 2004, 21(10): 809–12.
[44] VISCA P, SEIFERT H, TOWNER K J. Acinetobacter infection–an emerging threat to human health [J]. IUBMB Life, 2011, 63(12): 1048–54.
[45] IWAYA A, NAKAGAWA S, IWAKURA N, et al. Rapid and quantitative detection of blood serratia marcescens by a real-time PCR assay: its clinical application and evaluation in a mouse infection model [J]. FEMS Microbiology Letters, 2005, 248(2): 163–70.
[46] TZIROS G T. Alternaria alternata causes leaf spot and fruit rot on loquat (eriobotrya japonica) in Greece [J]. Australasian Plant Disease Notes, 2013, 8(1): 123–4.

Figures

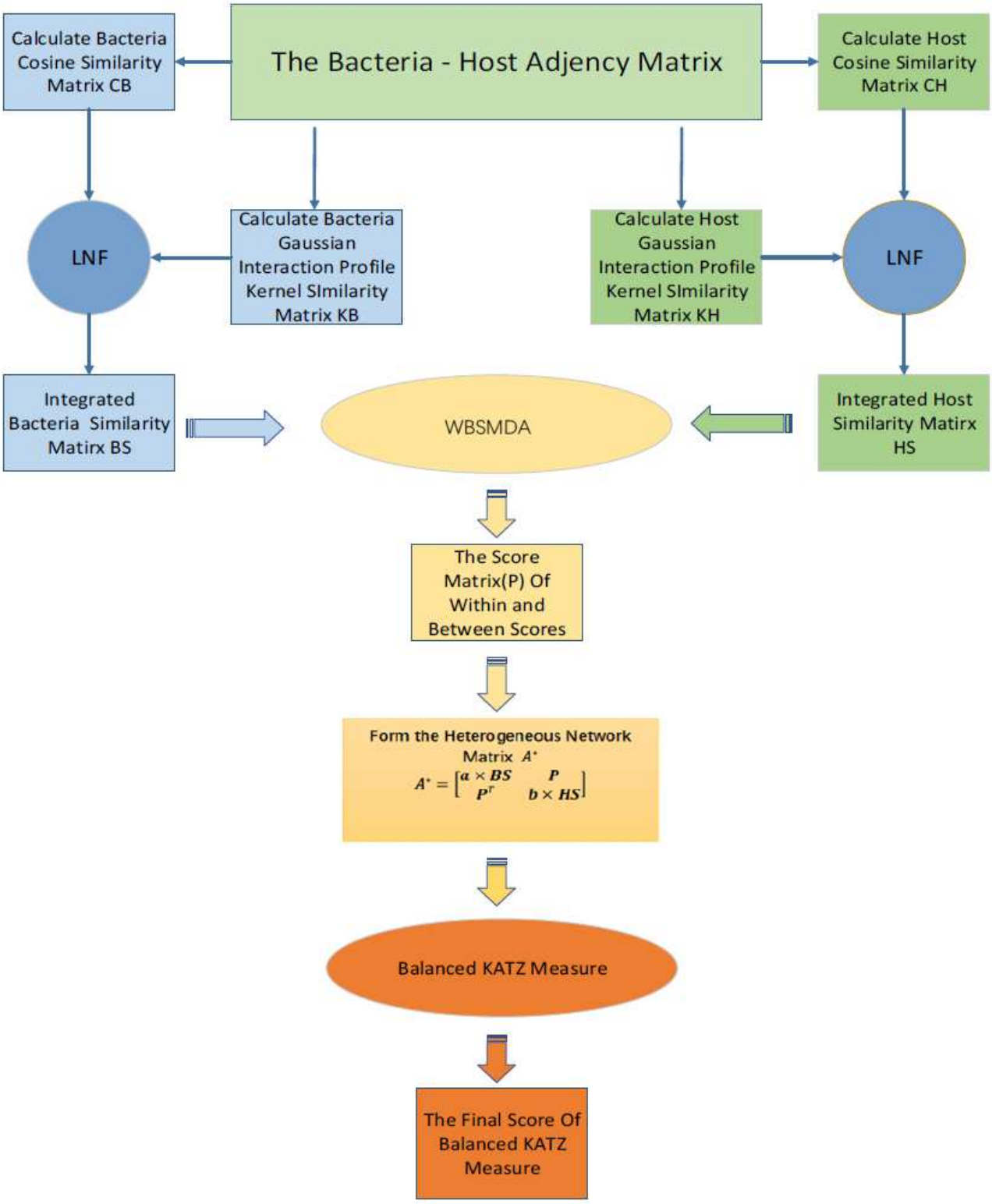

Figure 1

The process of ISKATZWBS

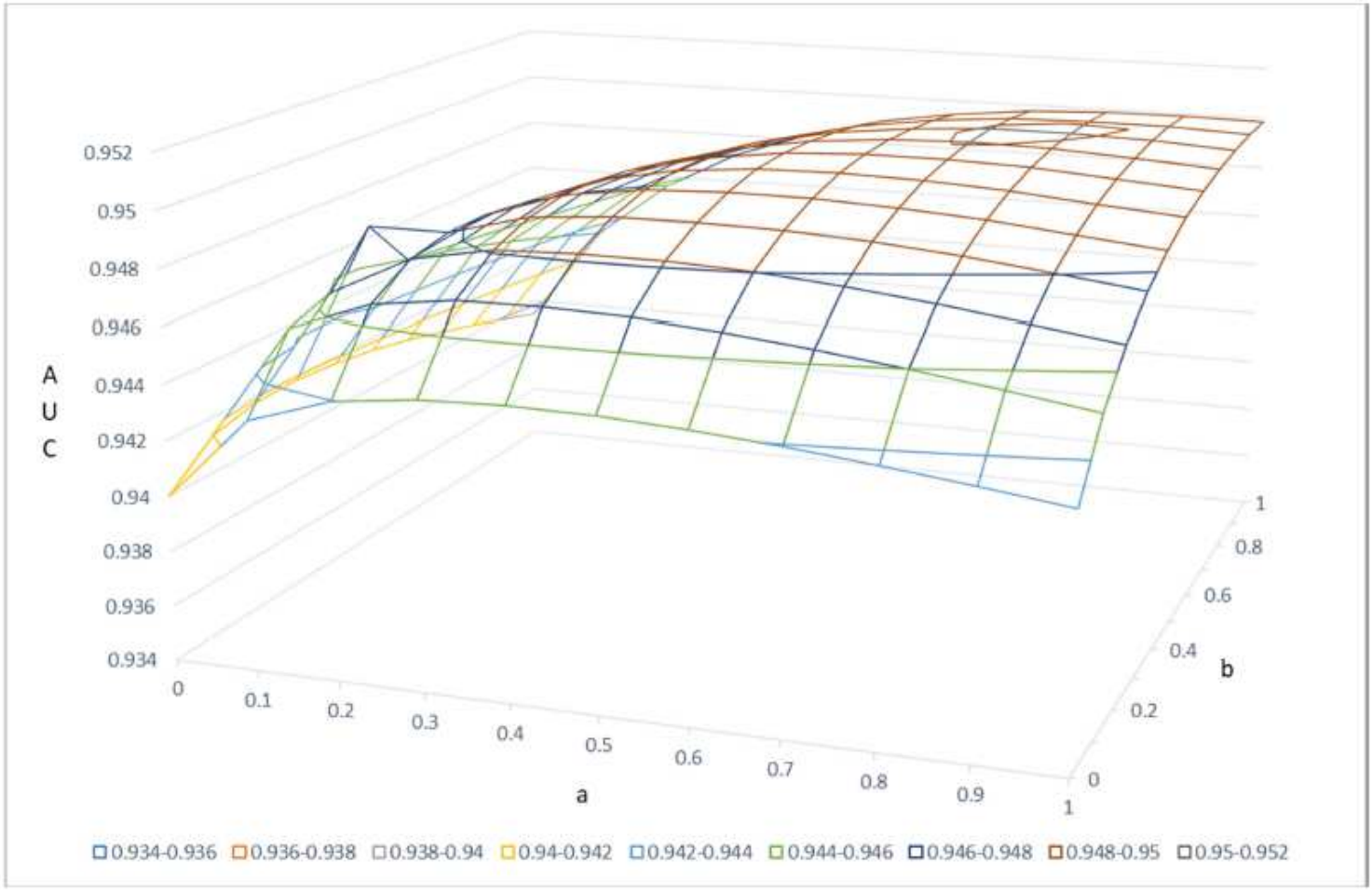

Figure 2

The LOOCV AUC results of ISKATZWBS with different parameters and

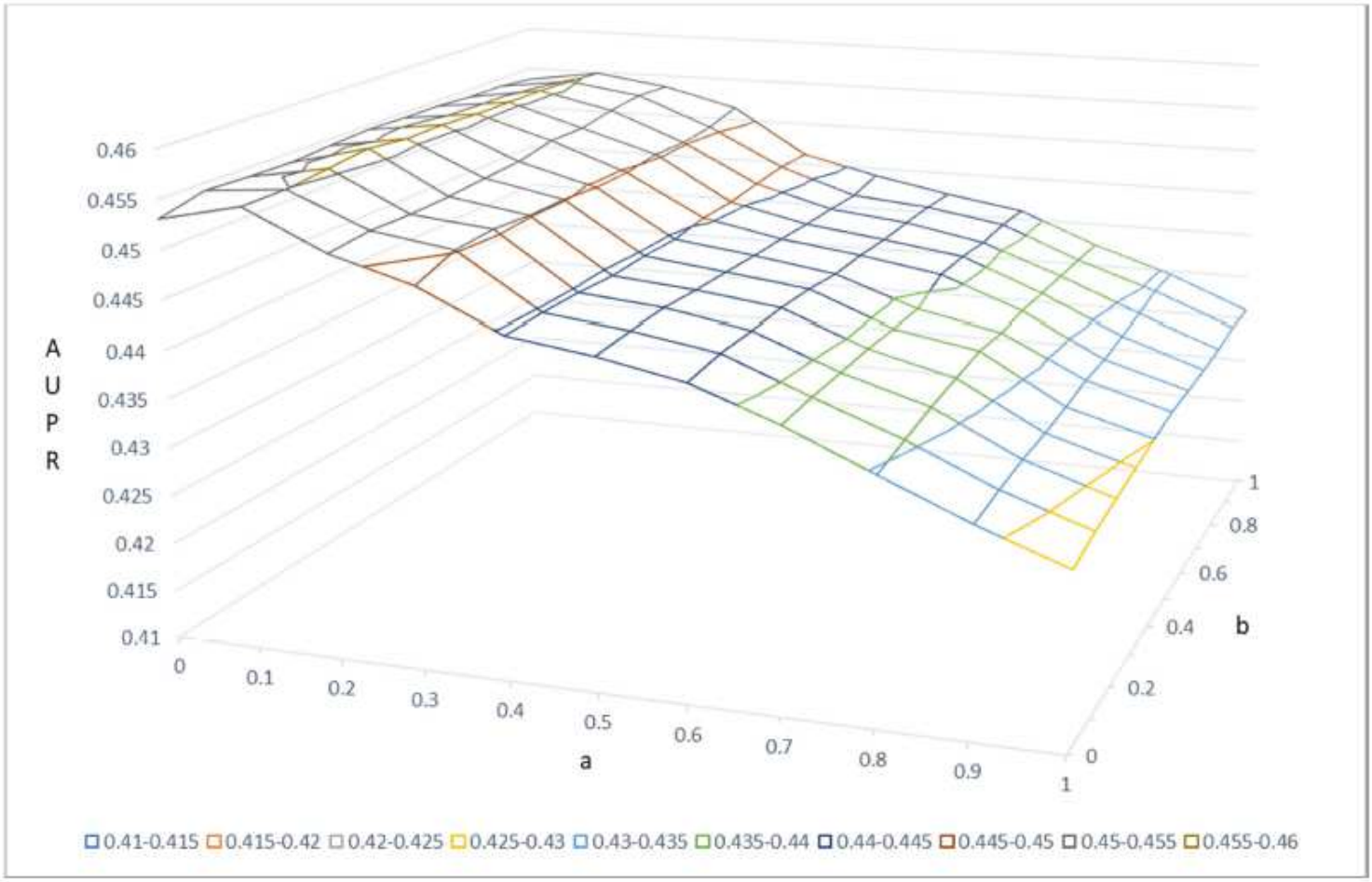

Figure 3

The LOOCV AUPR results of ISKATZWBS with different parameters and

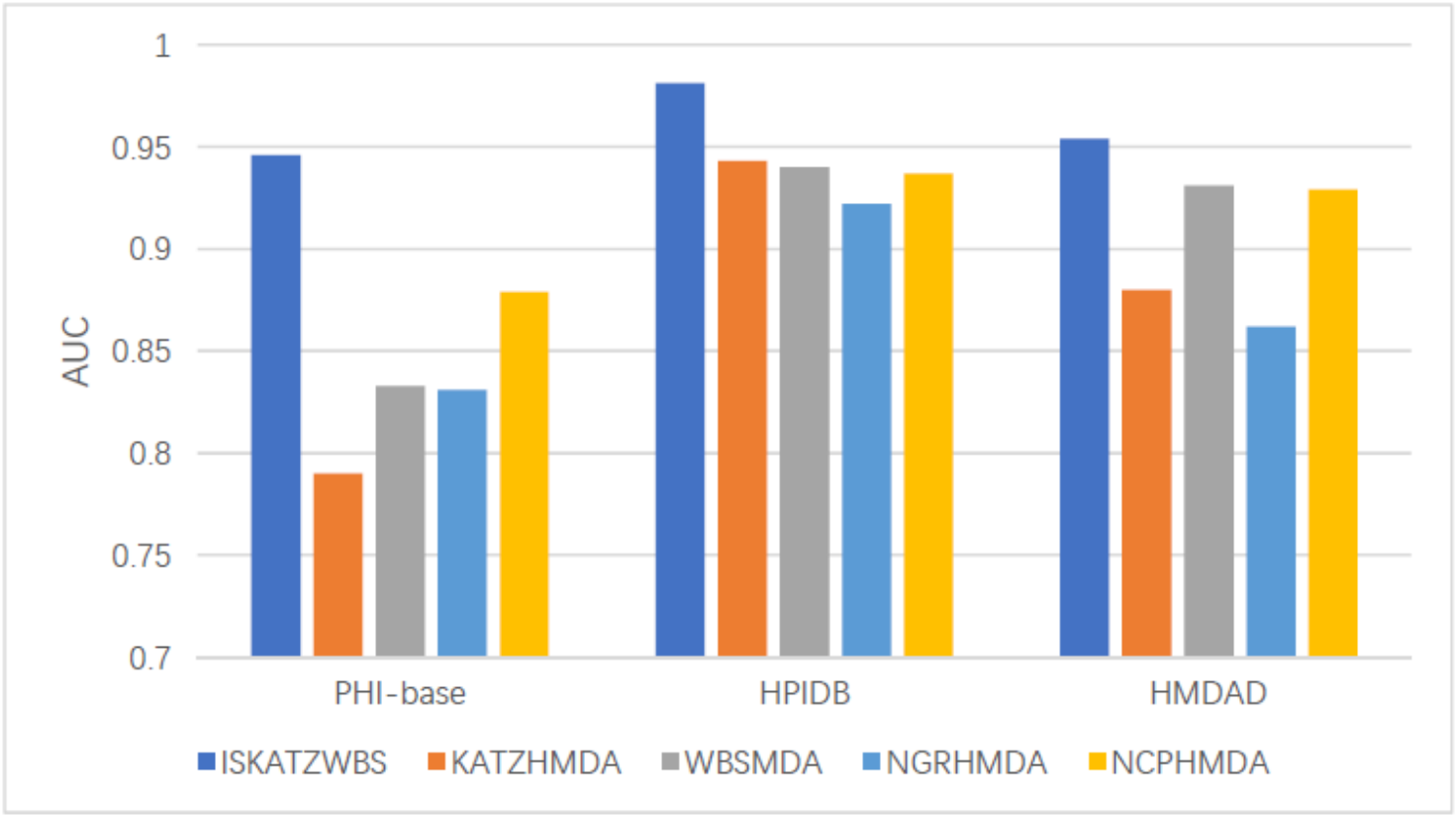

Figure 4

The AUC results of five methods on three data

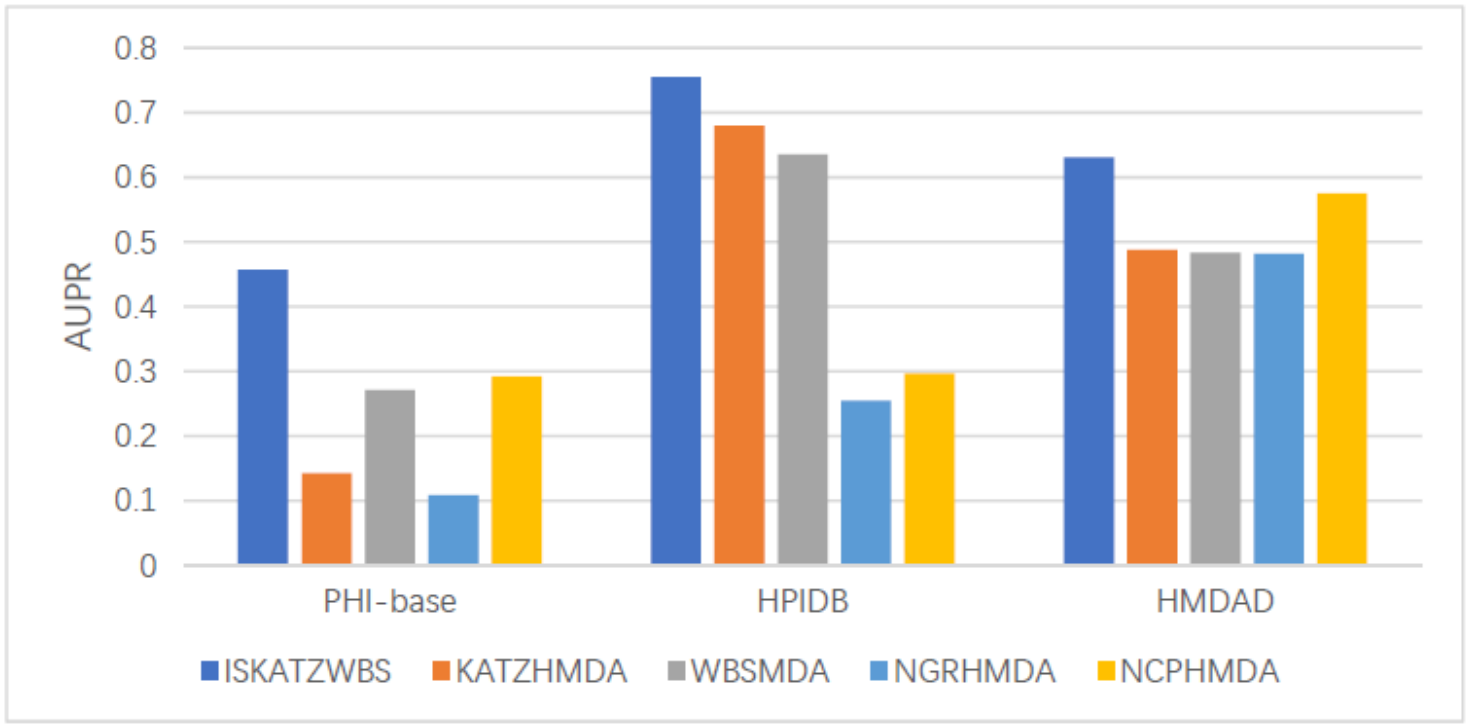

Figure 5

The AUPR results of five methods on three data