

Hypergraph Clustering Based on Game-Theory for Mining Microbial High-Order Interaction Module

Limin Yu^{1,2,3}, Xianjun Shen^{1,2,3} , Jincal Yang^{1,2,3}, Kaiping Wei^{1,2,3}, Duo Zhong^{1,2,3} and Ruilong Xiang^{1,2,3}

¹School of Computer, Central China Normal University, Wuhan, China. ²Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, Hubei, China. ³National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan, Hubei, China.

Evolutionary Bioinformatics
Volume 16: 1–8
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320970572



ABSTRACT: Microbial community is ubiquitous in nature, which has a great impact on the living environment and human health. All these effects of microbial communities on the environment and their hosts are often referred to as the functions of these communities, which depend largely on the composition of the communities. The study of microbial higher-order module can help us understand the dynamic development and evolution process of microbial community and explore community function. Considering that traditional clustering methods depend on the number of clusters or the influence of data that does not belong to any cluster, this paper proposes a hypergraph clustering algorithm based on game theory to mine the microbial high-order interaction module (HCGI), and the hypergraph clustering problem naturally turns into a clustering game problem, the partition of network modules is transformed into finding the critical point of evolutionary stability strategy (ESS). The experimental results show HCGI does not depend on the number of classes, and can get more conservative and better quality microbial clustering module, which provides reference for researchers and saves time and cost. The source code of HCGI in this paper can be downloaded from <https://github.com/yilm0505/HCGI>.

KEYWORDS: Microbial higher-order module, game-theory, hypergraph clustering, evolutionary stability strategy

RECEIVED: September 2, 2020. **ACCEPTED:** October 12, 2020.

TYPE: Machine Learning Models for Multi-omics Data Integration – Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China [61532008, 61872157, 6192008]; the Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE [CCNU19QD003]; and the National Language Commission Key Research Project [ZD1135-61].

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Xianjun Shen, Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, Hubei 430000, China. Email: xjshen@mail.ccnucnu.edu.cn

Introduction

There are many kinds of microorganisms, which exist in human, plant, soil, ocean and other environments. Microbial community can be defined as a collection of microorganisms that exist at the same time and may interact with each other. They exist in a certain spatiotemporal habitat. The term microbial community was first coined by Lederberg and McCray¹ By definition, microbial community is ubiquitous in nature, which has a huge impact on the living environment and human health.²⁻⁴ All these impacts of microbial community on the environment and its host are usually called the functions of these communities, which largely depend on the composition of the community, that is to say, on the species and quantity of the existing species. In essence, the interaction between species involves multiple species, which often occurs in a high-order combination way, that is, the interaction between two species is regulated by one or more other species,⁵ and the high-order interaction often dominates the function of microbial community, for example, higher-order functional interaction dominates the hydrolysis rate of amylase in biological communities.⁶

In order to quantify how the interaction between species affects the overall community function, two interactions between species were systematically measured through experiments, such as cross feeding between *E. coli* growth nutrient bacteria^{7,8} and naturally isolated bacteria.^{9,10} Other studies

infer the interaction between species by measuring the population dynamics of a complex community.¹¹⁻¹⁴ Some studies mainly focus on the interaction between two species and the ability of two interaction models to predict network functions.^{15,16} Although the interaction between the two is enough to describe the important impact of ecology in some cases, some studies have emphasized the need to merge higher-order relationships, mainly because of their role in population dynamics.^{5,17,18}

Tsai et al¹⁹ developed a rule-based microbial network (RMN) algorithm, which uses a triple relationship similar to the activation inhibitor target model to build a gene regulatory network. Based on this principle, RMN algorithm was developed to describe the relationship between every three microorganisms as a cooperation competition target. Guo and Boedicker²⁰ found that the interaction between species has an impact on the activity of the whole community, and studied the interaction network of four microbial communities through theoretical models. The results show that the interaction between multiple species determines the overall metabolic rate of the microbial community.²⁰ Bairey et al⁵ found that the high-order interaction has a lower bound to diversity, while the pairwise interaction exerts an upper bound. When the pairwise interaction is dominant and the number of species increases beyond the threshold, the feasibility decreases. On the contrary, when the four-way interaction is dominant, the



community becomes sensitive to species removal. In the case of mixing, when the pairwise interaction and the high-order interaction are dominant, the community becomes sensitive to species removal. When combined, the intermediate species is the most feasible, and the removal and addition of species will destroy the stability of the community. Therefore, the combination of high-order interaction and pairwise interaction determines the most stable range of diversity. These findings emphasize the importance of high-order species interaction in determining the diversity of natural ecosystems. In order to study the contribution of pairwise interaction and higher-order interaction to community function, Alicia Inspired by the complex genetic interaction, etc. Used the ecological function landscape to separate and quantify the single, paired and higher-order interactions in the microbial community function, screened the experimental results through the enzyme dynamics theory, and found that the higher-order functional interaction dominated the hydrolysis rate of amylase in our community.⁶ That is to say, it is very important to build a meaningful microbial network to study the high-order interaction between microorganisms.

From the perspective of modules, we can intuitively understand the interaction information and the contribution of network modules to the overall network, so as to study the possible functions of modules formed by specific microorganisms in the microbial community. Girvan and Newman²¹ proposed a method to detect the community structure, based on the centrality index to find the community boundary, and applied it to study many different social networks and biological networks, focusing on the least central edge and the most central edge between communities, not by adding the strongest edge to an initial empty vertex set to build the community, but by gradually removing edges from the original graph to build a community can help you understand the composition of other complex datasets.

Copeland et al²² used the network analysis method to study the relationship between the abundance of each genus in the leaf microbial community, determined two highly connected clusters, and found functional groups. Aleksej considering the possibility of high-order interaction involving more than two species, extended the concept of binary symbiosis to at most four species at the same time, and proposed a model for large microbial communities. The experimental results showed that although there was obvious resource competition at the level of the whole community, microbial communities still had interdependent metabolic groups, and they occur repeatedly in different habitats, emphasizing that metabolic dependence is the main driving force of species symbiosis, and suggesting that the cooperative group is a recurring module in the microbial community structure.²³ That is to say, the study of higher-order microbial modules can help us to understand the dynamic development and evolution process of microbial communities, help us to understand the functional composition of communities, and predict the potential functionality of community modules.

The traditional clustering problem only studies the pairing relationship between two microorganisms. Most current algorithms

assume that there are paired similarities between microorganisms, but many microorganisms in the real environment involve more complex and multipath relationships. In such networks of high order relationships, modeling pairs only results in the loss of information in the original data. The graph model of pairwise relations does not consider higher order information, and cannot represent the whole ecological network. Microorganisms diversity depends heavily on the stabilization of higher-order interactions. It is important to study the higher-order interaction of microorganisms to understand the diversity and stability of microbial community. Hypergraph clustering refers to the process of extracting the most consistent group from a group of objects by using high-order (rather than pairwise) similarity. Therefore, hypergraph clustering is applied to study the high-order interaction network module of microorganisms.

In this paper, first, the microbial abundance data were processed, from the perspective of logical relationship, the information entropy is used to determine the high-order interaction relationship of microorganisms, and then the high-order relationship hypergraph of microorganisms is constructed. Then, a kind of microbial logical relationship is selected for further study, and game theory is introduced to transform hypergraph clustering problem into clustering game problem. Finally, we use Baum-Eagon theorem to generate the clustering module with iterative elimination strategy until the remaining objects cannot form a high-order relationship.

Materials and Methods

The dataset

In this paper, the available microbial abundance data of healthy humans based on 16s rRNA comes from the human microbiome project (HMP), which is V13 high quality files processed with the mothur software package. The data covered 18 of the 5 main body parts. Data source web site is <http://hmpdacc.org/HMMCP/>.

Hypergraph clustering based on game theory

In the previous work, we used to find the maximum degree of modularity to determine the number of clusters before the premise of clustering.²⁴ In this paper, we studied the hypergraph clustering problem from the perspective of game theory to analyze the high-order module analysis of microorganisms, which can automatically generate clusters without specifying the number of clusters.

In the process of biological evolution, animal adaptability is formed in the interaction with their living environment. In the competition, animals finally choose the evolutionary stability strategy (ESS), which is adopted by most members of the population and will not be eroded by other strategies. In our framework, the clustering problem is seen as a non-cooperative game between multiple microorganisms, and an important concept in game theory is equilibrium, in which the performance of each player in a population does not

exceed the population average return when Nash equilibrium is reached. Buló et al²⁵ proved that the problem of finding these equilibrium points (clusters) is equivalent to solving a polynomial optimization problem with linear constraints, and used an algorithm based on Baum–Eagon inequality to solve this problem. We apply it to the analysis of the microbial high order module in this paper.

Game theory. A game can be represented as $\Gamma = (P, S, \pi)$. $P = \{1, \dots, k\}$ is the set of players in game, and $S = \{1, \dots, n\}$ is a set of pure strategies. π is a payoff function, Each player has their own set of strategies and payoff functions. The evolution of populations occurs because we hypothesize that there is a selection mechanism, similar to Darwin's evolutionary process, that spreads the fittest survival strategies in a population to the detriment of the weakest.

Given the set of all possible states of the population Δ :

$$\Delta = \left\{ x \in \mathbb{R}^n : \sum_{i \in S} x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i \in S \right\} \quad (1)$$

x_i represents the fraction of i -strategists in the population. If $y^{(i)} \in \Delta$ determine which strategy represents the i th player will adopt to play the game Γ . The average payoff obtained by the agents can be defined as:

$$u(y^{(1)}, \dots, y^{(k)}) = \sum_{(s_1, \dots, s_k) \in S^k} \pi(s_1, \dots, s_k) \prod_{j=1}^k y_{s_j}^{(j)} \quad (2)$$

This function insensitive to order of inputs. We suppose that randomly pick people from $x \in \Delta$ to play game Γ , the population average payoff is calculated by $u(X^k)$, X^k is a shortcut for a sequence (x, x, \dots, x) , and the average payoff that an i -strategist obtains in a population is calculated by $u(e^i, X^{k-1})$. $x \in \Delta$ is in equilibrium when the distribution of the strategy does not change any more, it is in an equilibrium state.

Indeed, $x \in \Delta$ is a Nash equilibrium if

$$u(e^i, X^{k-1}) \leq u(X^k), \text{ for all } i \in S. \quad (3)$$

Hypergraph clustering based on game theory. $H = (V, E, s)$ is a hypergraph clustering problem, V is a finite set of vertices, $V = \{1, \dots, n\}$ is a collection of microbes in various parts of the body, a node corresponds to a microorganism, which is also a set of objects to be clustered. E is a set of hyperedges, $s: E \rightarrow \mathbb{R}$ is a weight function that assigns a real value to a hyperedge, and it also is a similarity to the set of objects in E . And $\Gamma = (P, S, \pi)$ is the corresponding clustering game. The payoff function π intuitively does this by rewarding k players based on the similarity of the items they play. As a result, assuming that $(v_1, \dots, v_k) \in V^k$ is group of objects chosen by k players, the payoff function can be defined as:

$$\pi(v_1, \dots, v_k) = \begin{cases} \frac{1}{k!} s(\{v_1, \dots, v_k\}) & \text{if } \{v_1, \dots, v_k\} \in E, \\ 0 & \text{else,} \end{cases} \quad (4)$$

where $\frac{1}{k!}$ is a coefficient.

Given a population $x \in \Delta$ paly clustering game, we can obtain the average population payoff $u(X^k)$, which represents the average similarity of the objects that make up the cluster. The average payoff $u(e^i, X^{k-1})$ of $i \in V$ in population x represents the average similarity of object i with respect to the cluster.

The problem of extracting ESSs of our hypergraph clustering game can be transformed into the problem of finding strict local solutions of (5). Consider the following nonlinear optimization problem:

$$\text{maximize } f(x) = \sum_{e \in E} s(e) \prod_{i \in e} x_i \quad x \in \Delta \quad (5)$$

By the Karush-Kuhn-Tucker (KKT) conditions there exist $\mu_i \geq 0, i \in S, \lambda \in \mathbb{R}$ so that for all $i \in S$,

$$\begin{aligned} \nabla f(x)_i + \mu_i - \lambda \\ = \frac{1}{k} u(e^i, X^{k-1}) + \mu_i - \lambda = 0 \text{ and } \mu_i x_i = 0, \end{aligned} \quad (6)$$

∇ is gradient operator, we can see for all $i \in S$, $u(e^i, X^{k-1}) \leq u(X^k)$. In addition, it prove that any strict local maximizer $x \in \Delta$ of (5) is a ESS of Γ .

The hypergraph clustering problem of extracting ESSs (evolutionary stability) can be transformed into finding a strict solution of (5). Baum and Eagon²⁶ are introduced to find ESSs.²¹ For maximizing polynomial function in probability domain, The Baum-Eagon inequality is an effective iterative method.

Theorem 1 (Baum-Eagon). Defined $p(x)$ as a homogeneous polynomial in the variable x_i with non-negative coefficients, let $x \in \Delta$. Define the mapping $t = \Phi(x)$, we can obtain t_i as follow:









$$t_i = \frac{x_i \partial_i P(x)}{\sum_{j=1}^n x_j \partial_j P(x)}, i = 1, \dots, n \quad (7)$$

Then $P(\Phi) > P(x)$ unless $\Phi(x) = x$, Φ is the growth transformation of the polynomial P .

Using the above method, we iteratively find a cluster and remove it from the object set, then repeat the process on the rest of the objects until the remaining nodes cannot form a hyperedge after eliminating the nodes in the cluster.

The data processing. The traditional microbial network construction is based on graph-theory. In this paper, eight logical relationships (see Table 1) among microorganisms are calculated based on the method of entropy. Bowers et al²⁷ proposed a computational method based on genomic data to identify the detailed

Table 1. Description of the logical relationship between microbes and the total number of microbes present in the human body.

| TYPE | WAYNE FIGURE | LOGIC DESCRIPTION | TOTAL |
|-------|---|---|-------|
| type1 |  | C is present if and only if both A and B are present | 655 |
| type2 |  | C is present if A is absent or B is absent | 4 |
| type3 |  | C is present if A is present or B is present | 1890 |
| type4 |  | C is present if A is absent and B is absent | 1 |
| type5 |  | C is present if A is present (absent) and B is absent (present) | 354 |
| type6 |  | C is present if A is absent (present) or B is present (absent) | 93 |
| type7 |  | C is present if one of either A or B is present | 1334 |
| type8 |  | C is present if both A and B are present(absent) | 36 |

relationships between proteins. Based on information entropy, Bowers et al calculates correlations between proteins and identifies eight logical relationships, which reveals many previously unknown higher-order relationships. We applied it to construct microbial higher-order networks.

Table 1 describes logical relationships among microorganisms and the total number of each logical relationship in human body based on the selected dataset. Venn diagram and related logic statements in the Table 1 illustrate eight different logical relationships, which represent different co-occurrence relationships among three Microorganisms. In other words, these describe the possible dependence of the existence of C on the existence of A and B. It can be seen from Table 1 that the most common logical relationships among microorganisms in human body are type1, type3 and type7. The discussion of logical relations with less relations is of little significance. We can choose type1, type2 or type3, but the number of type 3 relationships is large, and the interaction between microorganisms is more complex, which is not easy to analyze, so type 1 is finally selected.

Figure 1 shows the calculation process of high-order logical relationship of microorganisms. $M(p, s)$ represents microbial abundance matrix. The element value $m(i, j)$ of $M(p, s)$ represents the abundance data of microbial i and sample j . The value of data that satisfies $m(i, j) > 0$ in $M(p, s)$ is set to 1, We can get a matrix with values of 0 and 1, then remove the weakly

expressed microorganisms, that is, remove the rows that meet

$\sum_j m(i, j) < 4$, we obtained the final microbial abundance matrix.

Select a logical relationship. We select type1. Based on type1, we calculate the uncertainty coefficient of $U(C|A)$, $U(C|B)$, $U(C|f(A, B))$ by formula (8), which predict microorganism C through two other microorganism A and B, and $f(A, B)$ is the logical combination and represents whether there is a logical relationship between microorganism A and microorganism B. Moreover, it is also required that neither microorganism a nor microorganism B can predict microorganism C independently. The maximum fraction value of $U(C|f(A, B))$ is obtained, and we select triplets that satisfy the uncertainty between two microorganisms described C is weak ($U(C|A) < 0.3, U(C|B) < 0.3$) and the uncertainty of describing C based on logical correlation is strong ($U(C|f(A, B)) > 0.5$).

$$U(X|Y) = \frac{[H(X) + H(Y) - H(X, Y)]}{H(X)} \quad (8)$$

$H(X)$ and $H(Y)$ in equation (8) denote the entropy of individual distribution, $H(X, Y)$ denotes the entropy of joint distributions. U is between 0 and 1, when U is 1, X is a deterministic function of Y , and when U is 0, X is completely independent of Y .

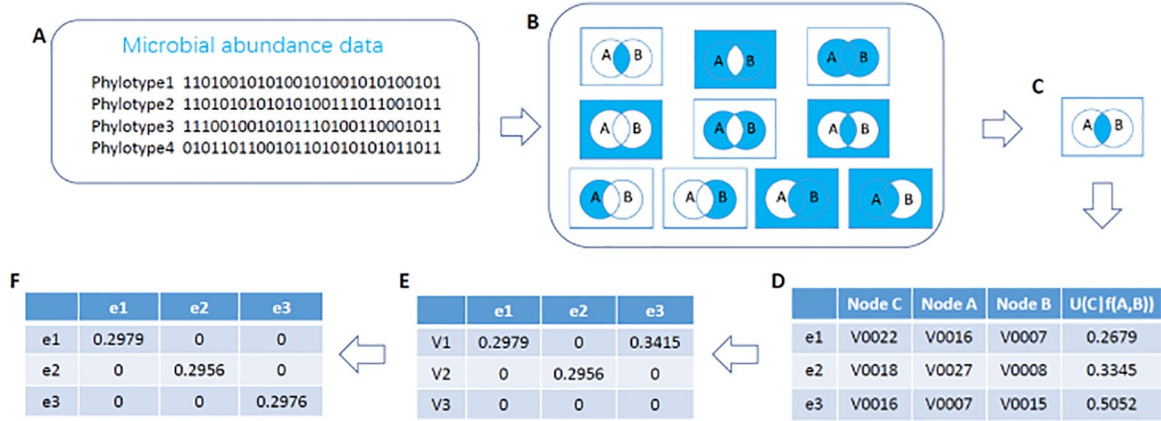


Figure 1. The calculation process of high-order logical relationship of microorganisms. (A) Extract microbial abundance data from open source websites converted into 0 to 1 microbial abundance matrix. (B) Calculate the number of 8 high order logical relationships in the presence of microorganisms. (C) Select type1. (D) Calculate the uncertainty coefficient $U(X|Y)$ as the initial hyperedge weight by entropy. (F) The final microbial hyperedge similarity matrix W_s is calculated through intra-class scatter matrix.

The maximum fraction value of uncertainty coefficient obtained by the triplet is taken as the initial weight of the hyperedge.

Construction of weight matrix $s(e)$. Weight each hyperedge with a positive value, we can defined it with $s(e)$. In our previous work, based on the idea of hypergraph clustering and the idea that the samples are as dense as possible, and the smaller the dispersion degree within the samples, the better the classification ability of the categories, we proposed that the method of reconstruct the microbial hyperedge similarity based on the intra-class scatter matrix to the analyze high-order microbial network modules.²⁰

$M = \{m_1, m_2, m_3, \dots, m_n\}$ denotes a collection of microorganisms, where m_i is dimensional vector, N_j is the number of samples j and M_j is a set of samples j . M can be consider as a relationship matrix of microorganisms nodes and hyperedges. The rows of the matrix represent the hyperedges, the column s of the matrix represent the microbes, if the microbes node j belongs to the hyperedge i , then the value of $M(i, j)$ is the initial weight of the hyperedge. μ_j is the mean vector of the hyperedge j , We define the intra-class scatter matrix as:

$$I = \sum_{m \in M_j} (m - \mu_j)(m - \mu_j)^T \quad (9)$$

$$\mu_j = \frac{1}{N_j} \sum_{m \in M_j} m.$$

Trace (\cdot) is the average measure of characteristic variance. The smaller trace (\cdot) is, the smaller the characteristic variance is and the higher the tightness within the class is. So, We can construct the weight of the hyperedge in the following way:

$$s(e_j) = 1 / \left(\text{trace} \left(\frac{e^I}{\delta} \right) \right) \quad (10)$$

Where e is the natural logarithm function, δ is a positive parameter, we can obtain the new hyperedge similarity matrix based on the scatter matrix $s(e)$.

Results and Analysis

Evaluation index

Joint entropy. The joint entropy (JE) is a measurement uncertainty evaluation method to measure the uncertainty related to a set of variables. Take JE as a standard of measuring the information contained in a cluster. Joint entropy is a measure of uncertainty associated with a set of variables, which is used to measure the amount of information contained in the same cluster of microorganisms. The smaller the joint entropy, the smaller the amount of expression information. X_1, X_2, \dots, X_n denotes microorganisms, $p(X_1, X_2, \dots, X_n)$ denotes is the probability that these microbes are present in the sample. The definition of JE can be expressed as:

$$JE = HX_1, X_2, \dots, X_n = - \sum_{x_1, x_2, \dots, x_n} p(x_1, x_2, \dots, x_n) \log p(x_1, x_2, \dots, x_n) \quad (11)$$

where HX_1, X_2, \dots, X_n is the amount of information that were transferred by the average value of random variables per batch $\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$.

Total correlation. Total correlation (TC)²⁸ is derived from mutual information generalization. Mutual information is an information measure in information theory, which can be regarded as the information contained by one random variable about another random variable. TC is an effective method to measure the independence between a set of microbes and to evaluate the degree of microbial interdependence within the same module. When the total correlation is close to 0, the microorganisms in the module are statistically independent. The lower the total correlation, the lower the interdependence,

Table 2. The clustering analysis and comparison of mid vagina based on type1.

| ALGORITHM | HCGI | | HCIS | |
|-----------|----------|----------|----------|----------|
| | JE | TC | JE | TC |
| 1 | 5.869237 | 1.726926 | 5.869237 | 1.726926 |
| 2 | 4.539703 | 4.502064 | 6.036468 | 6.45178 |
| 3 | 3.068427 | 1.553164 | 4.984123 | 8.21099 |
| 4 | 6.318696 | 5.685631 | 3.644253 | 2.288655 |
| 5 | 5.860217 | 4.757338 | 7.025616 | 12.27174 |
| 6 | 4.902228 | 0.78664 | – | – |
| 7 | 3.556522 | 0.433895 | – | – |
| Sum | 34.11503 | 19.44566 | 27.5597 | 30.95009 |

Table 3. The clustering analysis and comparison of intestines tract based on type1.

| ALGORITHM | HCGI | | HCIS | |
|-----------|----------|----------|----------|----------|
| | JE | TC | JE | TC |
| 1 | 6.219162 | 3.650304 | 4.793256 | 0.568318 |
| 2 | 5.492254 | 1.661277 | 6.219162 | 3.650304 |
| 3 | 5.114142 | 0.8542 | 6.439209 | 5.489505 |
| 4 | 8.542772 | 21.46201 | – | – |
| Sum | 25.36833 | 27.62779 | 17.45163 | 9.708126 |

thus ensuring the quality of the high-order interaction in the module.

$$\begin{aligned}
 & TC(X_1, X_2, \dots, X_n) \\
 &= \sum_{i=1}^n H(X_i) - H(X_1, X_2, \dots, X_n)
 \end{aligned} \quad (12)$$

Comparison of algorithms

In this paper, the joint entropy, total correlation and modularity are taken as the indexes to evaluate the clustering performance of the algorithm, the hyperspectral clustering based on game theory and intra-class scatter matrix (HCGI) and hypergraph clustering based on intra-class scatter matrix (HCIS).²⁴ They were applied to cluster analysis and comparison of microorganisms in 18 parts of human body. Based on the clustering results, the JE and TC of each cluster were calculated. we take the sum of JE of the subclusters and the sum of TC as an final index to evaluate the performance of an algorithm. The sum of JE represents the total expression information of the microbial cluster, and the sum of TC represents the total dependence within the microbial community.

We selected the mid vagina where the clustering effect did not change much after the addition of intra-class scatter matrix and intestines tract where HCIS did not work well in to compare and analyze in the previous work.²⁴

Tables 2 and 3 show that cluster analysis and comparison of mid vagina and intestines tract based on type 1. We can see that the two methods end up with different clustering numbers on the same body part. From Table 2, we can observe that the sum of JE generated by HCGI algorithm is higher than that of HCIS, while the sum of TC is lower than that of HCIS, which indicates that the results generated by HCGI clustering algorithm express more total information, and the correlation between the generated clusters is weaker. But in Table 3, the correlation between the generated clusters by HCIS is weaker, this is because the nodes removed by the two methods are not consistent, so it is not good to use the above evaluation indicators to evaluate the clustering performance of the two methods. However, according to joint entropy and total correlation of each cluster, we can judge the amount of expressed information of the cluster and the connection strength between nodes. Next, through the case analysis of mid vagina, we can visualize the clustering results of mid vagina, and specifically analyze the advantages and disadvantages of two methods.

Visual analysis

Cytoscape²⁹ has proven to be a high-level platform for visualization and analysis of biological networks. We used cytoscape to analyze the clustering results of HCGI on mid vagina and find the differences between nodes within each cluster.

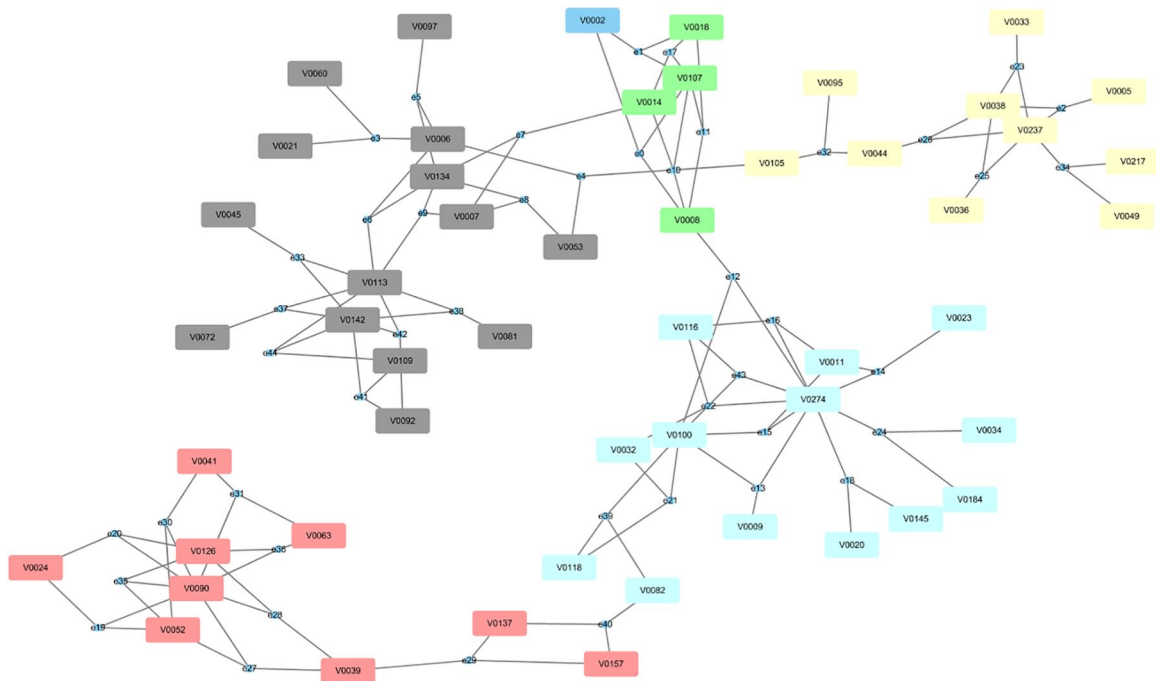


Figure 2. The cluster results of HCIS on mid vagina.

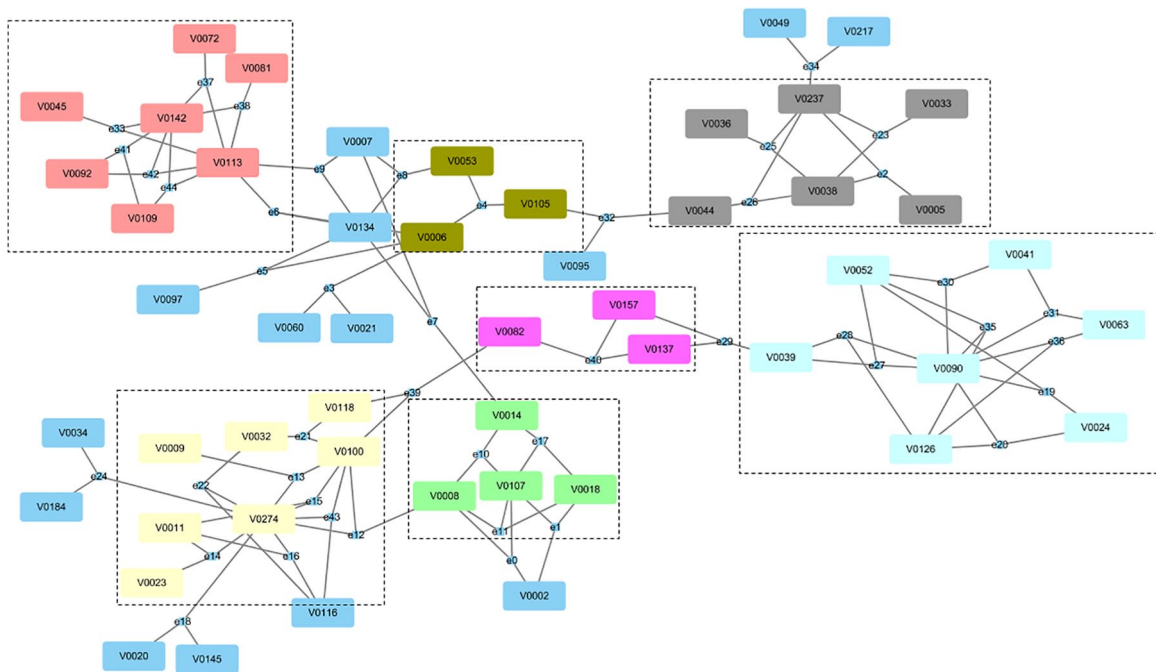


Figure 3. The cluster results of HCGI on mid vagina.

Figure 3 shows the clustering results of HCGI algorithm on mid vagina. The nodes in the box inside the black dotted box belong to the same cluster. HCGI algorithm divides the nodes on mid vagina into 7 clusters. The blue nodes that are not framed by the black dotted line do not belong to any of the clusters. Figure 2 shows the clustering results of HCIS on mid vagina. HCIS divides the nodes on mid vagina into 5 clusters. In Figure 2, the blue square nodes do not belong to any of the clusters. Compare the clustering results of Figures 2 and 3, we

can find that HCGI divides more clusters than HCIS algorithm. However, it can be observed that the cluster generated by HCGI algorithm is contained in the cluster generated by HCIS algorithm. The nodes eliminated by HCGI algorithm are also nodes connected sparsely to the network graph, and the completeness of the higher-order organization is still maintained in the process of iterative elimination. Based on the framework of game theory, we can get better and more conservative microbial modules.

Conclusion

Studying microbial community interactions is challenging because the microorganisms in a community often interact with each other in many ways, not just in pairs. The basic way to detect the biological significance of biological networks is based on the module perspective.

On the basis of hyperedge similarity matrix reconstructed by intra-class scatter matrix in our previous work, this paper introduces game theory and proposes hypergraph clustering based on game theory (HCGI) algorithm for mining microbial higher-order module research. Hypergraph clustering problem were converted to non-cooperative multi-player game clustering problems, and get the optimal clustering module by finding countermeasures balance. There is no need to specify the cluster number in advance, and one module is obtained after each iteration until the remaining object set cannot build the higher-order relationship. The experimental results show that the introduction of game theory makes the clustering module more conservative and the quality of the module better, which indicates that the higher-order module of hypergraph clustering based on game theory is effective.

Future Work

This paper is mainly based on information entropy to build non directional high-order microbial interaction relationship. The interaction relationship in real biological network may be directional relationship, such as metabolic relationship. Therefore, we can analyze and construct the possible high-order relationship between directed microorganisms from the perspective of metabolism.

With the development of time series data, many dynamic models have been used in microbial networks. Dynamic models such as dynamic Bayesian models can be used to construct high-order interaction networks among microorganisms.

Author Contributions

LY and XS designed and implemented the computing framework. LY and XS analyzed the results and wrote the manuscript. LY, XS, JY, KW, DZ and RX revised the manuscript. LY prepared the computational codes and carried out. All the authors wrote, reviewed and approved the final manuscript.

ORCID iD

Xianjun Shen  <https://orcid.org/0000-0001-5714-8848>

REFERENCES

- Lederberg J, McCray AT. Ome SweetOmic—a genealogical treasury of words. *Scientist*. 2001;15:22-27.
- Baker BJ, Jillian FB. Microbial communities in acid mine drainage. *Fems Microbiol Ecol*. 2003;44:139-152.
- Andersen R, Chapman SJ, Artz RRE. Microbial communities in natural and disturbed peatlands: a review. *Soil Biol Biochem*. 2013;57:979-994.
- Fuhrman JA. Microbial community structure and its functional implications. *Nature*. 2009;459:193-199.
- Bairey E, Kelsic ED, Kishony R. High-order species interactions shape ecosystem diversity. *Nat Commun*. 2016;7:12285.
- Sanchez-Gorostiaga A, Bajic D, Osborne ML, Poyatos JF, Sanchez A. High-order interactions dominate the functional landscape of microbial consortia. Preprint. Posted online May 29, 2018. bioRxiv 333534. doi:10.1101/333534
- Wintermute EH, Silver PA. Emergent cooperation in microbial metabolism. *Mol Syst Biol*. 2010;6:407.
- Mee MT, Collins JJ, Church GM, Wang HH. Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci USA*. 2014;111:E2149-E2156.
- Freilich S, Zarecki R, Eilam O, et al. Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun*. 2011;2:1-7.
- Vetsigian K, Jajoo R, Kishony R. Structure and evolution of *Streptomyces* interaction networks in soil and in silico. *PLoS Biol*. 2011;9:e1001184.
- Mounier J, Monnet C, Vallaers T, et al. Microbial interactions within a cheese microbial community. *Appl Environ Microbiol*. 2008;74:172-181.
- Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol*. 2014;5:219.
- Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One*. 2014;9:e102451.
- Needham DM, Fuhrman JA. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol*. 2016;1:1-7.
- Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol*. 2012;10:538-550.
- Stein RR, Bucci V, Toussaint NC, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol*. 2013;9:e1003388.
- Grilli J, Barabás G, Michalska-Smith MJ, Allesina S. Higher-order interactions stabilize dynamics in competitive network models. *Nature*. 2017;548:210-213.
- Goldford JE, Lu N, Bajic D, et al. Emergent simplicity in microbial community assembly. *Science*. 2018;361:469-474.
- Tsai K-N, Lin S-H, Liu W-C, Wang D. Inferring microbial interaction network from microbiome data using RMN algorithm. *BMC Syst Biol*. 2015;9:54.
- Guo X, Boedicker JQ. The contribution of High-Order metabolic interactions to the global activity of a four-species microbial community. *PLoS Comput Biol*. 2016;12:e1005079.
- Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci USA*. 2002;99:7821-7826.
- Copeland JK, Yuan L, Layeghifard M, Wang PW, Guttman DS. Seasonal community succession of the phyllosphere microbiome. *Mol Plant Microbe Interact*. 2015;28:274-285.
- Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc Natl Acad Sci USA*. 2015;112:6449-6454.
- Yu L, Shen X, Jiang X, Yang J, Yang Y, Zhong D. Hypergraph clustering based on intra-class scatter matrix for mining higher-order microbial module. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, 18-21 November, 2019.
- Bulò SR, Pelillo M. A game-theoretic approach to hypergraph clustering. *Adv Neural Inf Process Syst*. 2009;35:1571-1579.
- Baum LE, Eagon JA. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull Am Math Soc*. 1967;73:360-363.
- Bowers PM, Cokus SJ, Eisenberg D, Yeates TO. Use of logic relationships to decipher protein network organization. *Science*. 2004;306:2246-2249.
- Watanabe S. Information theoretical analysis of multivariate correlation. *IBM J Res Dev*. 1960;4:66-82.
- Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape automation: empowering workflow-based network analysis. *Genome Biol*. 2019;20:185.