Exploring the interaction patterns in seasonal marine microbial communities with network analysis

Shao-Wu Zhang*, Ze-Gang Wei, Chen Zhou, Yu-Chen Zhang, Ting-He Zhang
College of Automation
Nothwestern Polytechnical University
Xi'an 710072, China
Corresponding author: zhangsw@nwpu.edu.cn

***Abstract*—With the development of high-throughput and low-cost sequencing technology, a large amount of marine microbial sequences is generated. So, it is possible to research more uncultivated marine microbes. The interaction patterns of marine microbial species and marine microbial diversity are hidden in these large amount sequences. Understanding the interaction pattern and structure of marine microbe have a high potential for exploiting the marine resources. Yet, very few marine microbial interaction patterns are well characterized even with the weight of research effort presently devoted to this field. In this paper, based on the 16S rRNA tag pyrosequencing data taken monthly over 6 years at a temperate marine coastal sits in West English Channel, we employed the CROP unsupervised probabilistic Bayesian clustering algorithm to generate the operational taxonomic units (OTUs), and utilized the PCA-CMI algorithm to construct the spring, summer, fall, and winter seasonal marine microbial interaction networks. From the four seasonal microbial networks, we introduced a novel module detecting algorithm called as DIDE, by integrating the dense subgraph, edge clustering coefficient and local modularity, to detect the interaction pattern of marine microbe in four seasons. The analysis of network topological parameters shows that the four seasonal marine microbial interaction networks have characters of complex networks, and the topological structure difference among the four networks maybe caused by the seasonal environmental factors. The marine microbial interaction patterns detected by DIDE algorithm in four seasons show evidence of seasonally interaction pattern diversity. The interaction pattern diversity of fall and winter is more than that of spring and fall, which indicates that the seasonal variability might have the greatest influence on the marine microbe diversity.**

Keywords—marine microbe; operational taxonomic unit (OTU) ; interaction pattern; network; clustering

I. INTRODUCTION

Marine microbes are the 'gatekeepers' for the Earth System with an estimated contribution to global primary productive of between 50% and 90% [1], and also are the important composition in the marine ecosystem. Yet specific ecological relationships among these taxa are largely unknown. This is part due to the dilute, microscopic nature of the planktonic microbial community, which prevents direct observation of their interactions [2]. With the development of high-throughput DNA sequencing technologies that yield a mass of reads of small-subunit rRNA gene (16S rRNA/18S rRNA) and DNA, we can describe the compositions of microbial communities, their diversity and how communities may change across space, time or experimental treatments based on these sequence data [3]. However, most of the current analytical approaches of describing and comparing the structure of communities often focus on the total numbers of taxa, the relative abundances of individual taxa and the extent of phylogenetic or taxonomic overlap between communities or community categories[4-6]. In contrast, there has been far less attention focused on using sequence data to explore the direct or indirect interactions between microbial taxa coexisting in environmental samples. Although some researchers used the network analysis to explore co-occurrence pattern in soil and ocean[2-3, 7-9], they just constructed the associate networks to show the co-occurrence pattern, and did not further mine the networks finding the pattern structures. The microbial interaction (or co-occurrence) patterns can offer new insight into the structure of complex microbial communities, reveal the niche spaces shared by community members, identify habitat affinities or shared physiologies that could guide more experimental settings. In this article, we will construct the spring, summer, fall and winter marine microbial interaction networks, introducing a new module detecting algorithm to find the microbial interaction patterns. The aim was to understand the relationship among microbe and seasonal variability and try to determine the microbial interaction pattern difference among seasons.

II. MATERIAL AND METHODS

A. Dataset

The 16S rRNA sequence dataset used in this paper were downloaded from http://vamps.mbl.edu/index.php, which includes 969,400 sequences generated from 76 time point seawater samples at the surface of L4 sampling site in the West English Channel [8]. The 76 seawater samples were arranged into spring (March-May), summer (June-August), fall (September-November) and winter (December-January) seasons, in which 22, 27, 13 and 14 samples belong to spring, summer, fall and winter seasons respectively. And the 16S rRNA sequence numbers of spring, summer, fall and winter seasons are 249,395, 293,549, 202,356 and 224,100, respectively. In order to establish the seasonal marine microbial network at the taxonomic level (e.g. species, genus), the 16S

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

rRNA sequences were grouped into species-level operational taxonomic units (OTUs) with CROP program [10], which resulted in 5858 OTUs. CROP adopts an unsupervised probabilistic Bayesian clustering algorithm and uses a soft threshold for defining the OTUs, which bypasses setting a subjective hard cut-off threshold, thus may effectively reduce the effects of PCR and sequencing errors in inferring OTUs.

B. OTU-OTU correlation network modeling

In order to research the correlation among different microbial species and explore their interaction patterns, we use a vector X_i to represent each OTU in the four seasons.

$$X_i = [x_{i1}, x_{i2}, \cdots, x_{is}, \cdots, x_{iS}] \quad (i=1, \ldots, 5858) \tag{1}$$

where x_{is} is the *i*-th OTU abundance value in the *s*-th sampling, that is, x_{is} equals the ratio of the sequence number N_{is} contained in the *i*-th OTU and the total sequence number N_s contained in the *s*-th sampling. To reduce the sequencing effort bias, the x_{is} value was set to zero if $N_{is} < 5$. For reducing the false higher correlation between vectors, we also remove these OTU vectors which contain less than 3 non-zero elements. After these processing, we can obtain 825 OTU vectors, in which spring season contains 205, summer 179, fall 208 and winter 233 OTUs respectively. Then, the four microbial abundance matrixes of spring, autumn, fall and winter season were constructed by normalizing every OTU vector.

The path consistency algorithm based on conditional mutual information (PCA-CMI) was firstly used to infer (or reconstruct) the gene regulatory networks (GRNs) from gene expression data [11]. In PCA-CMI algorithm, the conditional independence between a pair of genes is represented by conditional mutual information between this gene-pair given certain other genes. PCA-CMI can cover nonlinear relations between gene pairs based on mutual information (MI) and CMI from information theory, and detect the nonlinear statistical dependence. If the seasonal microbial abundance matrixes were considered as the gene expression data, we can use PCA-CMI to construct the seasonal marine microbial correlation networks. The process of PCA-CMI can be described simply as follows. Firstly, generate a complete graph G according to the number of OTUs. Secondly, compute the zero-order MI $I(i, j)$ for all OTU pairs.

$$I(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

If the OTU_i and OTU_j has low or zero MI, then the edge between OTU_i and OTU_j is deleted. Thirdly, for adjacent OTU_i and OTU_j, select the adjacent OTU_k of them and compute first-order CMI $I(i, j \mid k)$.

$$I(X,Y \mid Z) = -\sum_{x \in X, y \in Y, z \in Z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \tag{3}$$

If the OTU_i and OTU_j has low or zero CMI which represents independent correlation, then the edge between them is deleted. The next step is to compute higher order CMI until there are no more adjacent edges. In this paper, considering the computational complexity, we just compute the second-order CMI and construct the four second-order marine microbial correlation networks for spring, summer, fall and winter seasons.

C. DIDE network module detecting algorithm

Before introducing the DIDE algorithm, we will present some notions. 1) Number of shared adjacent nodes $|N_{ij}|$. For an undirected graph $G(V,E)$, suppose N_i and N_j are the adjacent node sets of node v_i and v_j respectively, the number of shared adjacent nodes between any two nodes is defined as $|N_{ij}| = |N_i \cap N_j|$. 2) Dense subgraph. Suppose the degree of node v_a is maximum, $N = \{v_1, v_2, \cdots, v_r\}$ is adjacent node set of v_a, $N' = \{v_1, v_2, \cdots, v_s\}$ is the adjacent node set of v_a shared the most nodes with *N*, $N'' = \{v_1, v_2, \cdots, v_t\}$ is the adjacent node set of v_a shared the second most nodes with *N*, then the dense subgraph is defined as $C_d = \{v_b \in (\{v_a\} \cup N' \cup N'')\}$. 3) Edge clustering coefficient C_{ij}. Suppose that there is an edge e_{ij} between node v_i and v_j, and the two nodes have a shared node v_l, and the edges e_{ij}, e_{il} and e_{jl} can form a triangle loop, then the edge clustering coefficient C_{ij} of edge e_{ij} is defined as $C_{ij} = (t_{ij}+1)/\min(d_i - 1, d_j - 1)$, d_i and d_j are the degree of node *i* and *j* respectively, t_{ij} is the total number of true triangle loops including the edge e_{ij}, and the denominator represents the total number of theoretical triangle loop including the edge e_{ij}. The DIDE algorithm can be described in detail as follows.

i) Select the maximum degree node forming the initial dense subgraph C_d, then producing the module *M*.

ii) Compute the module *M* local modularity value $Q_M = L_{in}/(L_{in} + L_{out})$, L_{in} is the inner edge number of the module *M*, L_{out} is the outer edge number of the module *M*.

iii) Search the adjacent node v_i of the module *M*; compute the connection tightness U_i between node v_i with the module *M*, $U_i = |E_{iM}|/d_i$, $|E_{iM}|$ means the connection edge number of node v_i with the module *M*, d_i is the degree of node v_i.

iv) If $U_i \geq \theta$, then add the node v_i to the module *M*, forming a new module *M'*.

v) Else, compute the edge clustering coefficient of other residue adjacent node v_j. If the edge clustering coefficient of node v_j with the module *M* is the biggest, then add the node v_j to the module *M*, forming a new module *M'*.

vi) Compute the *M'* local modularity value $Q_{M'}$, If $Q_{M'} - Q_M < 0$, then remove the node v_j from *M'*.

vii) Repeat iii) to vi) for all other adjacent node of the *M'*,

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

until $\Delta Q_M = 0$.

viii) If there is still other dense subgraph, return to i).

The flow diagram of DIDE algorithm is displayed in Fig.1.

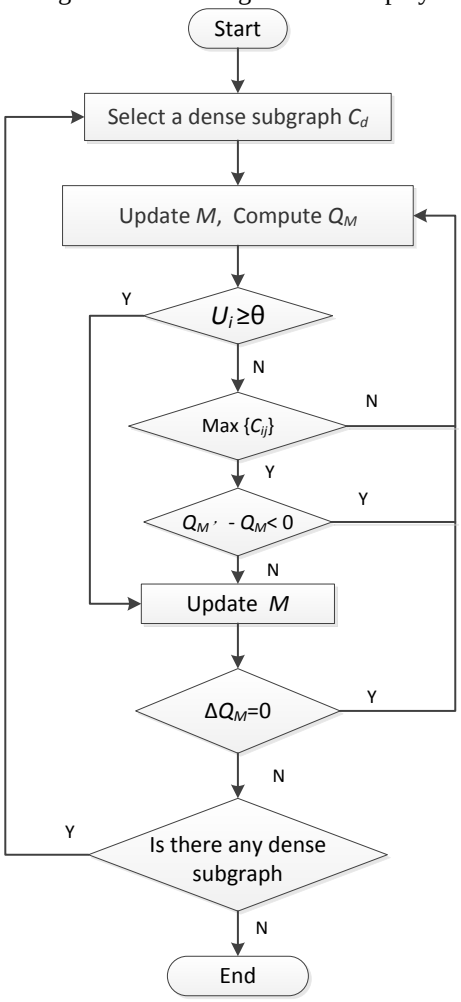
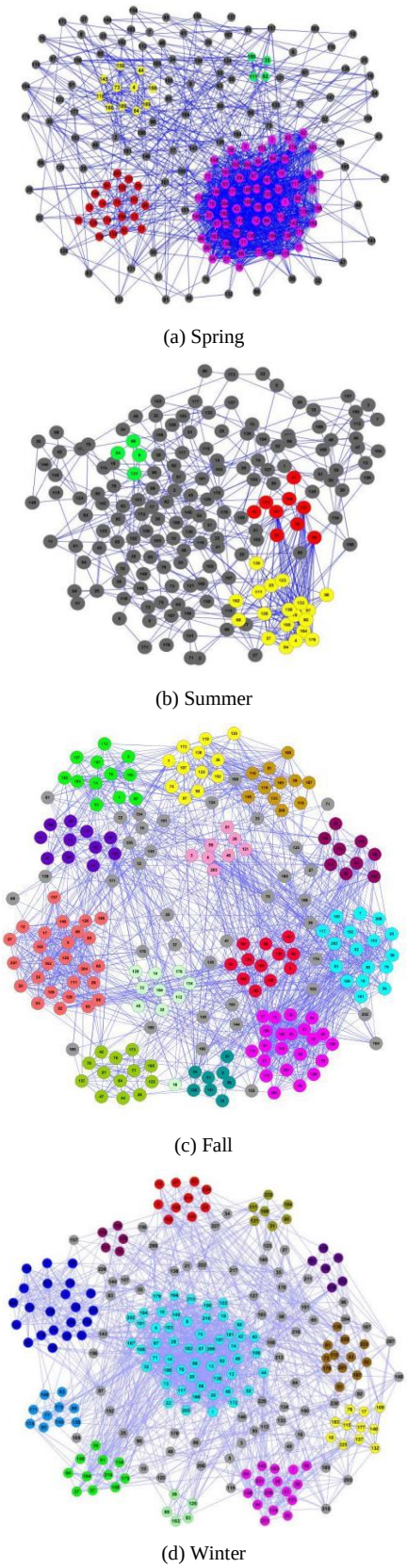


Fig. 1. The flow diagram of the DIDE algorithm

III. RESULTS AND DISCUSSION

A. *Topology analysis of four seasonal marine microbial correlation networks*

In order to study and analyze the microbial diversity of spring, summer, fall and winter seasons, we constructed the four seasonal marine microbial correlation networks with PCA-CMI algorithm. Fig.2 shows the four marine microbial correlation networks in spring, summer, fall and winter seasons. We also computed their topological parameters including the average degree, average clustering coefficient, average power law degree, modularity, and compared with their corresponding random networks. The comparison results of four seasonal networks and random networks are summarized in Table 1.



(a) Spring

(b) Summer

(c) Fall

(d) Winter

Fig. 2. Marine microbial correlation networks in spring, summer, fall and winter seasons.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

TABLE 1. TOPOLOGICAL PARAMETERS OF FOUR SEASONAL MARINE MICROBIAL CORRELATION NETWORKS AND THE CORRESPONDING RANDOM NETWORKS

	Seasonal networks				Random networks			
	Spring	**Summer**	**Fall**	**Winter**	**1**	**2**	**3**	**4**
Node Number	205	179	208	233	205	179	208	233
Edge Number	1419	512	1690	1791	1419	512	1690	1799
Avg. degree	13.8	5.8	16.3	15.4	13.8	5.8	16.3	15.4
Avg. clustering coefficient	0.37	0.20	0.43	0.343	0.065	0.015	0.073	0.065
Avg. power law degree	0.637	0.948	0.322	0.882	0.355	0.507	0.06	0.245
Modularity	0.461	0.571	0.553	0.493	0.223	0.377	0.212	0.219

From Table 1, we can see that there is some difference on the topological parameters among the spring, summer, fall and winter seasonal microbial correlation networks, which maybe caused by the seasonal environmental factors. Compared with random networks, four seasonal microbial correlation networks have bigger average clustering coefficient, average power law degree and modularity, which indicate that the four seasonal microbial correlation networks have some characters of complex network.

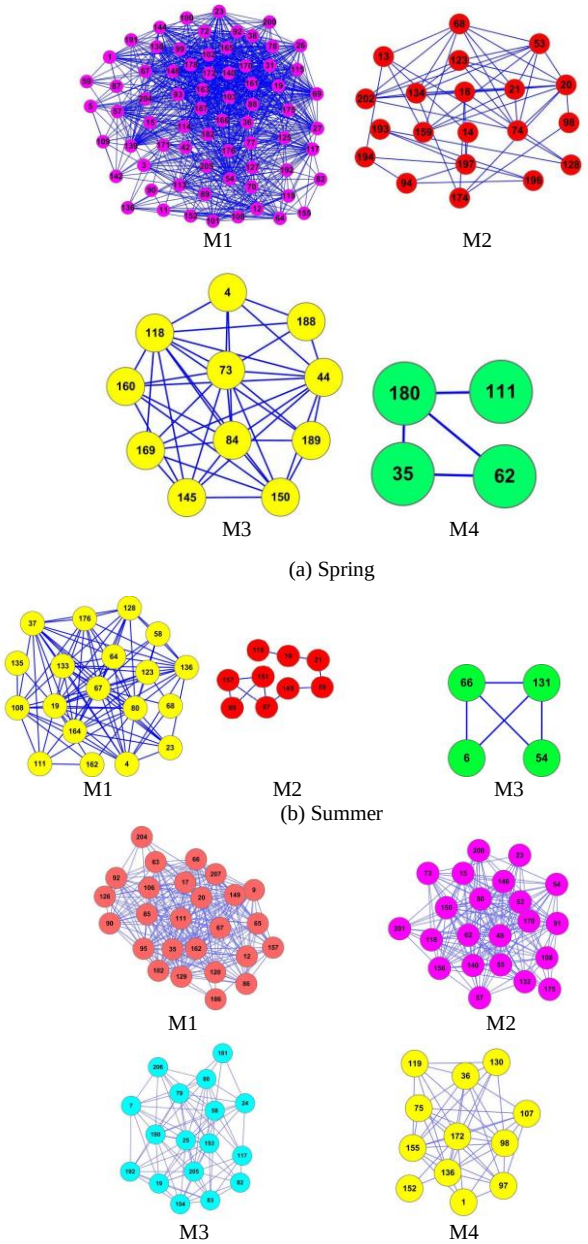
B. The interaction patterns in seasonal microbial networks detected by DIDE Algorithm

We firstly choose the typical American College Football network to verify the performance of DIDE algorithm. American Football Network contains 115 nodes and 613 edges [12]. Each node represents one team which belongs to one of the 12 Unions. Edge indicates the games played by the teams against other during the regular season of fall 2000. The performance of DIDE is better than that of Zhang's method [13], FN[14] and GN [15]. DIDE algorithm with U_i=0.5 can detect 12 communities, and the accuracy arrives at 91%. Yet Zhang's method, FN and GN just detect 10, 10, and 11 communities respectively, and their corresponding accuracies are 90%, 65% and 78% respectively. We also investigate the parameter U_i how to affect the results for four seasonal marine microbial OTU-OTU networks. The numbers of microbial interaction pattern detected by DIDE algorithm for spring, summer, fall and winter microbial OTU-OTU networks are listed in Table 2, and the structures of pattern at U_i=0.5 are show in Fig. 3.

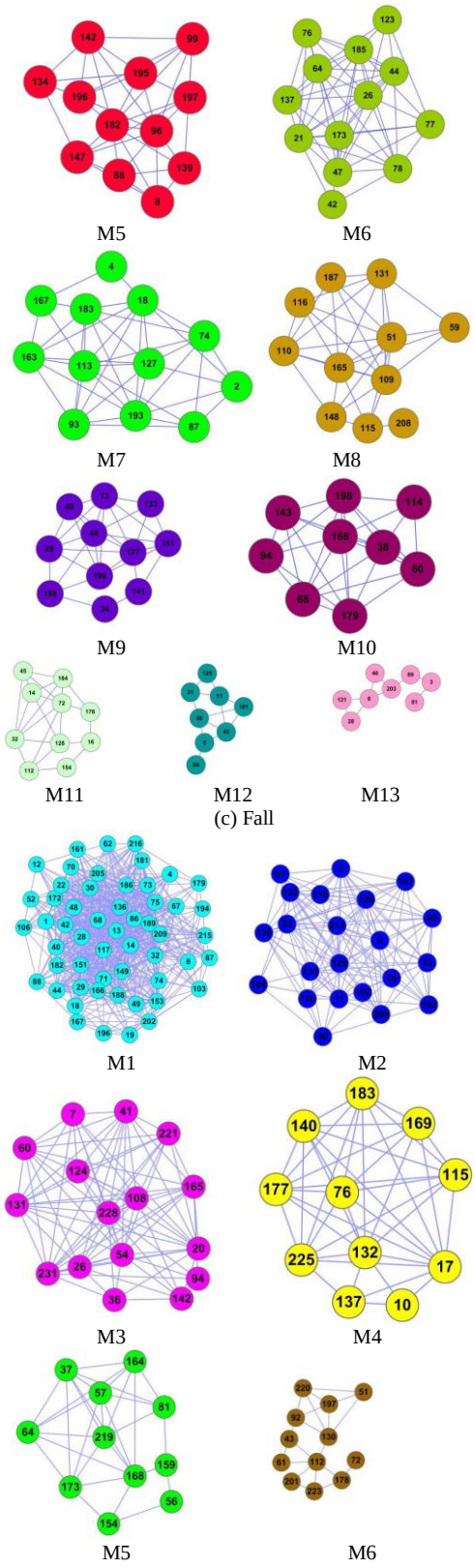
The results in Table 2 show that the interaction pattern diversity of fall and winter is more than that of spring and fall, which indicates that the seasonal variability might have the greatest influence on the marine microbe diversity, and the parameter U_i has little influence to the community number detected for U_i>0.5.

TABLE 2. NUMBERS OF MICROBIAL INTERACTION PATTERN DETECTED WITH DIFFERENT PARAMETER U_i

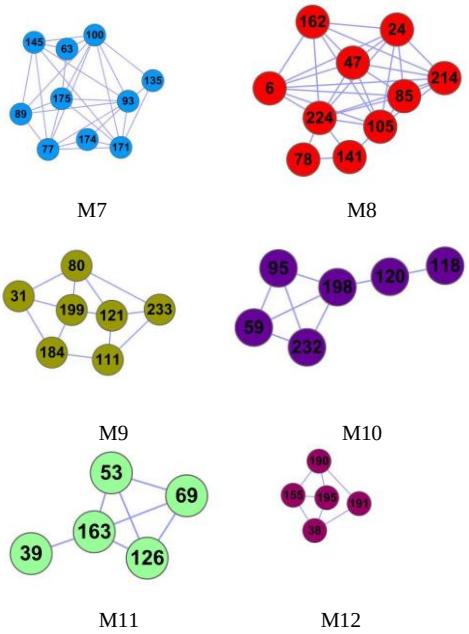
	0.4	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
spring	4	4	5	6	6	6	6
summer	2	3	3	2	2	2	2
fall	10	13	15	15	18	18	17
winter	1	12	13	18	20	20	20



978-1-4799-1389-3/13/$31.00 ©2013 IEEE



(c) Fall



(d) Winter

Fig. 3. The structure of microbial interaction pattern detected by DIDE algorithm at U_i=0.5

Fig. 3 shows that the interaction pattern diversity of fall and winter is more than that of spring and fall. According to the annotation information of OTUs at taxonomic level by using a number of different annotation strategies (e.g. GAST[4], BLAST against Greengenses[16], SIVA[17], RDP[18]), we analyzed in detail the biggest module for every seasonal network. The M1 module in spring microbial network has 71 OTUs in which the 44 OTUs were identified in Phylum level as 'Proteobacteria' , 7 OTUs as 'Actinobacteria', 5 OTUs as 'Bacteroidetes', 5 OTUs as 'Deferribacteres', 4 OTUs as Chloroplast, 1 OTU as 'Gemmatimonadtes', 1 OTU as 'Verrucomicrobia' and 4 OTUs could not be identified; in family level, the 7 OTUs were identified as 'Rhodospirillaceae', 6 OTUs as 'Rhodobacteraceae', 5 OTUs as "SAR11", 5 OTUs as 'SAR406', 4 OTUs as 'SAR86', 3 OTUs as 'Cryomorphaceae', 2 OTUs as 'Comamonadaceae', 2 OTUs as 'Iamiaceae' , the other OTUs belong to 'Flammeovirgaceae', 'Falvobacteriaceae', 'Oceanospirillaceae', 'Piscirickettsiaceae', 'Propionibacteriaceae', 'Pseudomonadaceae', 'Salinisphaeraceae', 'Sinobacteraceae', 'Sphingomonadaceae', 'Subdivision3' respectively, and 22 OTUs could not identified. Due to more 70% OTTUs of M1 in spring network could not be identified to genus, we did not further analyze M1 structure under genus taxonomic level.

The M1 module in summer microbial network has 19 OTUs in which the 12 OTUs in Phylum level belong to 'Proteobacteria', 3 OTUs to 'Bacteroidetes' , 2 OTUs to 'Deferribacteres', the other OTUs belong to 'Cyanobacteria' and 'Verrucomicrobia' respectively.

The M1 module in fall microbial network has 26 OTUs in which the 12 OTUs in Phylum level belong to 'Bacteroidetes', 9 OTUs to 'Proteobacteria', 3 OTUs to

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

'Verrucomicrobia', the other two OTUs belong to 'Actinobacteria', 'Cyanobacteria' respectively.

The M1 module in winter microbial network has 54 OTUs in which the 38 OTUs in Phylum level belong to 'Proteobacteria', 10 OTUs to 'Bacteroidetes', 2 OTUs to 'Actinobacteria', the other two OTUs belong to 'Deferribacteres', 'Firmicutes' respectively, and residual two OTUs could not identified.

The M1 structural analysis in four seasonal microbial networks show that a large fraction microbial interactions in Phylum level occur among 'Proteobacteria', 'Bacteroidetes', 'Actinobacteria', especially OTUs from the same phyla (e.g. Proteobacteria) tended to co-occur more.

IV. CONCLUSIONS

Mining the marine microbial interaction patterns and diversity is a key for exploiting the marine resources. Considering the marine microbes are symbiosis or competition, exhibiting a numerous, significant intra- or inters- lineage associations, we used the network approach to analyze the potential interaction patterns among the marine microbes from the 16S rRNA sequences. The results show that the four seasonal marine microbial interaction networks have characters of complex networks, and the marine microbial interaction patterns are related with the seasonal variability. Although we cannot claim that we have a comprehensive view of interactions within marine microbial communities, our network analysis method is more feasible and interesting for exploring the unseen patterns emerged in the complex dataset, including non-random association, deterministic processes at different taxonomic levels and expected relationship between community members.

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China (No. No.61170134 and No. 60775012)

REFERENCES

[1] P.G. Falkowski, R.T.Barber, and V.Smetacek, "Biogeochemical controls and feedbacks on ocean primary production," Science, vol. 281, pp.200-206, 1998.

[2] J.A. Steele, P.D. Countway, L. Xia, P.D. Vigil, et al., "Marine bacterial, archaeal and protistan association networks reveal ecological linkages," The ISME Journal, vol. 5, pp.1414-1425, 2011.

[3] A.Barberan, S.T. Bates, E.O.Casamayor and N.Fierer, "Using network analysis to explore co-occurrence aptterns in soil microbial communities," The ISME Journal, vol.6, pp.343-351, 2012.

[4] M.L. Sogin, H.G.Morrison, J.A.Huber, et al., "Microbial diversity in the deep sea and the underexplored 'rare biosphere'," Proc. Natl. Acad. Sci. USA, vol.103, pp.12115-12120, 2006.

[5] J.A.Gilbert, D.Field, P.Swift, et al., "The taxonmoic and functional diversity of microbes at a temperate coastal site: A'Multi-Omic' study of seasonal and diel temporal variation," PLos One, vol.5, e15545, 2010.

[6] D.L. Kirchman, M.T. Cottrell and C. Lovejoy, "The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes," Environmental Microbiology, vol. 12, pp.1132-1143, 2010.

[7] J. Zhou, Y. Deng, F. Luo, Z. He and Y.Yang, "Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO2," mBio, vol.2, pp. e0122-11, 2011.

[8] J.A. Gilbert, J.A. Steele, J.G. Caporaso, L. Steinbruck, J.Reeder, B. Temperton, et al., "Defining seasonal marine microbial community dynamics," The ISME Journal, vol. 6, pp. 298-308, 2012.

[9] A. Eiler, F. Heinrich and S. Bertilsson, "Coherent dynamics and association networks among lake bacterioplankton taxa," The ISME Journal,vol.6, pp.330–342, 2012.

[10] X. Hao, R. Jiang, T.Chen, "Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering," Bioinformatics, vol.27, pp.611–618, 2011.

[11] X. Zhang, X.M. Zhao, K. He, L. Lu, et al., "Inferring gene regulatory networks from gene expression data by PC-algorithm based on conditional mutual information," Bioinformatics, vol. 28, pp. 98-104, 2012.

[12] M.Girvan, and M.E.J. Newman, "Network of American football games between division IA colleges during regular season fall 2000," Proc. Natl. Acad. Sci. USA, vol. 99, pp. 7821-7826, 2002.

[13] D.W. Zhang, F.D. Xie, Y. Zhang, et al., "Fuzzy analysis of community detection in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 389, pp.5319-5327, 2010.

[14] M.E.J. Newman, "Fast algorithm for detecting community structure in networks," Phys. Rev. E, vol.69, p. 066133, 2004.

[15] M.E.J. Newman, M.Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, vol.69, p.02611, 2004.

[16] T.Z.DeSantis, P.Hugenholtz, N.Larsen, et al., "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," Apple Environ Microbiol, vol.72, pp. 5069-5072, 2006.

[17] E. Pruesse, C.Quast, K. Knitttel, B.M. Fuchs, et al., "SILVA: a comparehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," Nucleic Acids Res., vol.35, pp.7188-7196, 2007.

[18] B.L. Maidak, J.R.Cole, T.G.Lilburn, et al., "The RDP-II (Ribosoma Database Project)," Nucleic Acids Res., vol. 29, pp.173-174, 2001.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE