A Similarity Network Approach for Analyzing the Marine Microbial Diversity

Wei Chen, Yongmei Cheng, Shaowu Zhang, Liyang Hao, Peng Ding College of Automation, Northwestern Polytechnical University, 710072, Xi'an, China Zhangsw@nwpu.edu.cn;chenwei903@gmail.com

Abstract-The microbes in the world's oceans are most abundant organisms on earth, playing an important role in the maintenance the balance of marine ecology. However, little knowledge of ecological interdependencies is known due to the limitation of current method for large-scale data and narrow surveys done for marine microbes while microbe exhibited significant inter-lineage associations naturally. Here we present a similarity network-based method to represent and analyze potential interactions among the marine microbes based on the 16S rRNA sequences. A set of parameters such as network degrees, short path, clustering coefficient and so on, are computed to characterize the similarity network topology. A few core sub networks (or network motifs) were found which show that microbe in the marine environment has a cluster propensity and evolutionary relatedness, meanwhile, the variable of network motif also indicated that the microbial diversity has a regional difference. These results show the network-based methods are effective for advance understanding the complexity and function of the marine microbial community after experiment technical.

Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Microbes have a key effect on biogeochemical cycle and have a closely association with human life [1]. In the early days, the studies on microbes were mainly based on the pure cultivate methods which lead to a little knowledge of microbes was found due to the limitation of cultivateindependent for more than 99% of species can not be cultivated [2]. Recent advance in metagenomics, the great progress in high-throughput technical such as massively parallel pyrosequencing, especially the invention of deep sequencing technique for rapidly and large-scale generating 16S rRNA sequence data, supplying more potential data resources for the analysis of microbes which have revealed large of microbial and functional diversity of microorganisms which predominate in the oceans, soils, polar bear, gut and other environments [1, 3-9].

Marine microbes, as the most part of the microbes on earth, hold for most of the ocean biomass which have an important role in maintaining the fine ecosystem balance, and are responsible for a largely potential source of primary products [10].With the development of the microbiology and the deep understanding of marine, more and more attentions were paied to the study on microbial and functional diversity. A far more accurate and complete picture of the community and activities for the marine microbes are provided and a number of diversity of previously unknown, uncultivated microscopic life forms in the ocean are beginning to be elucidated, such as the microbial ecology of four coral atoll [8], diversity of eukaryotic lineages [4], rare biosphere of marine bacteria [11] and seasonal structure of microbial communities at WCO [12]. These literatures on the marine microbes, having greatly expanding our understanding of marine microbes, however, are mostly focused on microbial diversity, or the relation between microbes and environmental factors. So far, what has not been addressed but more important is whether and to what extent marine microbes may interact with other species via specific communication.

During the last decade, some researches have showed that many biochemical activities such as nutrient uptake [13], cellular differentiation [14], as well as coral reef degraded [8] are involved by many species, and much bacterial are known to eavesdrop and to even respond to signals that they can not themselves generate isolated [15]. Meanwhile, the rapidly accumulating knowledge on intra- and inters- species microbial communication haves also showed that there maybe exist in a special manner which is different from commonly assumed among the microbes ecosystem. Unfortunately, apart from these, the knowledge of specific interactions or the associations among the microbes is still quite poor. For the current work, considering the interactions or associations involved on inter or intra species, we present a similarity network-based method, for the first time to our best knowledge, to study the marine microbial diversity and potential associations among different marine microbes.

II. METHODS

A. Dataset

The data used for our work were based on Mitchell' work [11], we choose three samples, the sample ID are 53R, 137 and 138. All of the trimmed 16S rRNA sequences which have been showed to contain sufficient information for the estimating the species diversity were downloaded from the GOLD database [16] and the number of trimmed tags per sample ranged from 12999 to 13804 sequences. In order to establish the similarity network for the marine microbes at the species-level, firstly, we group the 16S rRNA sequences in to operational taxonomic units (OTUs) based on the ESPRIT program [17] for per sample sets. The details of the samples are summarized in the TABLE I.

²⁰¹¹ IEEE International Conference on Systems Biology (ISB) 978-1-4577-1666-9/11/\$26.00 ©2011 IEEE

TABLE I. DATA SUMMARY AND SIMILARITY-BASED OTUS

Sample ID	Site	Trimmed tags	Unique tags	OTU
53R	Labrador seawater	12999	2102	1519
137	Labrador seawater	13804	1841	1370
138	Labrador seawater	13193	1935	1355

The OTUs were defined at 97% sequence identity by using the program ESPRIT.

B. Network Modeling

The marine microbes similarity network was constructed using each of the OTU (or microbial lineage) as the vertex in the network, and whether there is an edge occurred between two given OTUs based on the 16S rRNA sequence identity. The Mothur tool [18] was used to measure sequence similarity in order to model the similarity network and the arguments were set as following: 'match=1, mismatch=-1, gapopen=-2, gapextend=-1', which have been showed to be the best parameters for alignments for 16S rRNA gene sequences. When we obtained the similarity measurement for each pair of the sequences, we can define the similarity for pair of OUTs, which can be written as following:

$$Sim(OTU_{i}, OTU_{j}) = \sum_{k_{i}, k_{j}=1}^{|OTU_{i}|, |OTU_{j}|} S_{k_{i}, k_{j}} / |OTU_{i}| * |OTU_{j}| ; i, j = 1, 2, ...N$$
(1)

Where N is the total number of OTU per sample, $|OTU_i|$ is the total sequence number in the i-th OTU, S_{k_i,k_j} is the similarity for the k_i -th sequence in the OTU_i and k_j -th sequence in the OTU_i.

Based on the similarity for the pairs of OTUs, an undirected graph G= (V, E) was used to represent potential association between OTUs for each sample, where V represents the set of OTUs and E is set of the edges between the OTUs. It can be fully expressed by its adjacency matrix $A= (A_{i, j})$ where $A_{i, j}=1$ if there is an edge between OTU_i and OTU_j and the values for elements in $A_{i, j}$ are defined as following:

$$A_{i,j} = \begin{cases} 1 & \text{if } Sim(OTU_i, OTU_j) >= Threshold \\ 0 & otherwise \end{cases}$$
(2)

Where *Threshold* is a threshold and in this work we set *Threshold* =0.9. According the definition above, three marine microbes similarity network as established showed in Figure 1.

III. RESULT AND DISCUSSION

A. Topology of marine microbes network

In order study and analyze the difference among the different samples, a few global and local parameters were extracted for the constructed marine microbes network. The views from recent studies show that these statistics are remarkably constant across many biological networks, which plays an important role in understanding the network architecture and performance. The computed topology parameters included average clustering coefficient, average

2011 IEEE International Conference on Systems Biology (ISB) 978-1-4577-1666-9/11/\$26.00 ©2011 IEEE

)

network diameter, network density, betweenness and degree, etc [19]. The results for the three networks are summarized in TABLE II.

TABLE II. TOPOLOGY OF THE SIMILARITY NETWORK

Sample ID	53R	137	138
Avg. clustering coefficient (c)	0.534	0.473	0.469
Avg. connected components	64	71	70
Avg. network Diameter (d)	6	6	5
Number of nodes	308	291	295
Number of edges	1038	709	940
Network density	0.022	0.017	0.021
Short paths	5306(5%)	3292(3%)	3778(4%)
Avg. number of neighbors	6.740	4.873	6.373

In fact, we found that the topologies change considerably among the three similarity network. Though the nodes in the threes network are nearly the same, but the number of edges among the three network are quite different. Especially for the network of sample 53R, it has the highest average clustering coefficient, average number of neighbors and network density, showing that the similarity network of sample 53R is much denser than the other network and have more potential associations between the species. Furthermore, the network of sample 138 has the minimal value for the average network diameter while the parameter is used to measure the short path length over all pairs of nodes in the network, which shows the sample 138 has the fewest intermediate and may imply faster propagation of the signal and regulation. Meanwhile, the network of sample 137 has the highest connected components, which suggests a weaker connectivity and there maybe exist a less activity and interactions among the species.

B. Network communities in the marine microbes networks

Algorithms for finding network communities (or modules) were introduced to help one better understand the interplay between network structures and function because network motifs or communities have been uncovered in many types of social and biological networks, and have been proved to be functioned as essential components of the network. In this work, DPclus [20] was used to find network communities which have been proved to be efficient in determining network community with the biological network. And the communities for the three networks are showed in Figure 1.

From Fig. 1(a), it can show that twenty communities of size>=3 marked by the red circle were found when the clustering is performed using the parameter d_{in} =0.7 and cp_{in} =0.5, twenty-eight communities were found in the Fig. 1(b) for the sample 137 while two of them with a size>8 were detected, and twenty-three communities were detected for the sample 138 but only three of them with the size>8. Among the three network, the sample 53R has more big communities (size>8) than that of others. These results show that the architectures of network for the three samples are quite

different, indicating that there is an obviously difference in marine microbial communities and diversity, the community structure correlates with changes in area. In the sample 137, there maybe more species have a function relation or phylogenetically relation than that of sample 138 and sample 53R. In order to further explain some intriguing results above, quality of larger clusters with size>=8 obtained are evaluated with taxonomy, here we take the clustering results of the network of Sample 53R for example, and the detailed analysis results are listed in TABLE III.

From the TABLE III, we can see that similar species commonly cluster together, for example, the communities 1, 3 and 4 have the same class attribution that belong to rickettsiales, indicating there is a obviously parasitism or coexisting in the marine microbes. Meanwhile, though the taxonomy unit for species in the communities at class level is not certain, it still give us a new way to study the species and these species in the communities maybe share a common characterizes.



(a). DPclus cluster results for the microbes similarity network of the sample 53R



(b). DPclus cluster results for the microbes similarity network of the sample



(c). DPclus cluster results for the microbes similarity network of the sample 138

Figure 1. Undirected similarity network and clustering descriptions for the three samples

Communities	Vertex	taxonomy
1	OTU1175 OTU1157 OTU1172 OTU1213 OTU1170 OTU1154 OTU1163 OTU1196 OTU1193 OTU1129 OTU1516 OTU1202	Rickettsiales
2	OTU102 OTU1268 OTU1271 OTU1314 OTU1317 OTU1367 OTU1368 OTU1372 OTU1383 OTU1387 OTU105 OTU112 OTU1270 OTU1309 OTU1362 OTU1366 OTU1373 OTU1379 OTU1388 OTU1409 OTU1410 OTU1411 OTU1420 OTU1440	unclassified
3	OTU1197 OTU1150 OTU1190 OTU1220 OTU1173 OTU1124 OTU1156 OTU1208 OTU1258	Rickettsiales
4	OTU1123 OTU1162 OTU1152 OTU1133 OTU166 OTU1055 OTU1518 OTU1138 OTU1224 OTU113 OTU39 OTU94 OTU1184	Rickettsiales
5	OTU1495 OTU1100 OTU1499 OTU1484 OTU1094 OTU1099 OTU1064 OTU1078 OTU1485 OTU1102 OTU585 OTU602	unclassified
6	OTU1187 OTU1218 OTU1146 OTU1219 OTU1137 OTU429 OTU414 OTU421 OTU1149	unclassified
7 -	OTU1155 OTU1144 OTU1233	Rickettsiales
	OTU1139 OTU1126 OTU1211 OTU1145 OTU1143	unclassified

 TABLE III.
 The analysis results of communities with size>=8 in the similarity network of Sample 53R

IV. CONCLUSION

The marine microbes have vital effect on ecosystem balance and have a close relationship with human life. Considering the marine microbes in the ocean are symbiosis or competition, exhibiting a numerous, significant intra- or inters- lineage associations, a network-based method was first used to represent and analyze potential interactions among the marine microbes based on the 16S rRNA sequences. From the results, we can found that there is an obviously difference among the established similarity network, indicating there maybe have different eco-architectures for the three samples.

2011 IEEE International Conference on Systems Biology (ISB) 978-1-4577-1666-9/11/ $$26.00\$

Meanwhile, the results from the clustering show that the species from the same family are apt to cluster together, indicating the species which phylogenetically closely related occur more intra or inter-lineage associations. Above results also show that our network-based methods are effective and promising for researching the complexity and stability of the marine microbial community. Investigating function variability and function annotation is the next logical step, this will help to understand microbial diversity and ultimately function.

ACKNOWLEDGMENT

This paper was supported in part by the National Natural Science Foundation of China (No. 60775012 and 60634030), the Technological Innovation Foundation of Northwestern Polytechnical University (No.KC02), and Doctorate Foundation of Northwestern Polytechnical University (No. cx201017).

REFERENCES

- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The Human Microbiome Project. Nature 449: 804-810
- [2] Venter JC, Remington K. Environmental genome shotgun sequencing of the Sargasso Sea. Science, 2004, 6(1):19-34
- [3] Pace NR, Stahl DA, Lane DJ, Olsen GJ. (1986) The Analysis of Natural Microbial-Populations by Ribosomal-RNA Sequences. Advances in Microbial Ecology 9: 1-55.
- [4] Wei Luo, Huirong Li, Minghong Cai, Jianfeng He. (2009) Diversity of microbial eukaryotes in Kongsfjorden, Svalbard, Hydrobiologia, 636:233-248
- [5] Samuel C, Hubert R, Jakob P, et al. (2010) A global network of coexisting microbes form environmental and whole-genome sequence data, Genome Res, 2010. 20: 947-959
- [6] Trine Glad, Pål Bernhardsen, et al.(2010) Bacterial diversity in faces from polar bear (Ursus aritimus) in Arctic Svalbard, BMC Microbiology 2010, 10:10

- [7] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557
- [8] Elizabeth AD, Olga P, et al. (2008) Microbial Ecology of Four Coral Atoll in the Northern Line Island. PLos one,3(2):e1584
- [9] Naqvi A, Rangwala H, Keshavarzian A, Gillevet P.(2010) Networkbased modeling of the human gut microbiome.chem biodivers,7(5):1040-1050
- [10] Atlas RM, Bartha R. (1993) Microbial Ecology: Fundamentals and Applications (Benjamin/Cummings, Redwood City, CA).
- [11] Mitchell L. Sogin, Hilary G. Morrison, et al. (2006) microbial diversity in the deep sea and the underexplored rare biosphere, PNAS,103(32): 12115-12120
- [12] Jack A. Gilbert, Dawn Field, et al.(2009) The seasonal structure of microbial communities in the Western English Channel, Environ Microbiol. 11(12):3132-9
- [13] Kolter R, Greenberg E P. (2006) Microbial sciences: the superficial life of microbes. Nature 441(7091):300-302.
- [14] Waters C. and Bassler B. (2005) Quorum sensing: cell-to-cell communication in bacteria. Annual review of cell and developmental biology.21:319-346.
- [15] Visick KL, Fuqua, C. 2005. Decoding Microbial chatter: cell-cell communication in bacteria. J Bacterial, 187(16): 5507-5519.
- [16] Kongtantinos Liolios, I-Min A.Chen, et al.(2009) The Genomes On Line Database (GOLD) in 2009:status of genomic and metagenomic projects and their associated metadata, Nucleic Acids Research,38
- [17] Yijun Sun, Yunpeng Cai, Li Liu, et al.(2009) ESPRIT: Estimating Species Richness Using Large Collections of 16S rRNA Pyrosequences, Nucleic Acids Research,2009,1-8
- [18] Schloss PD, et al. Introducing mothur: Open-source, platformindependent, community-supported software for describing and comparing microbial communities. Appl Environ Microbial, 2009. 75(23):7537-41
- [19] Shimbel, A. (1953) Structural parameters of communication networks. Bull Math Biophys 15:501-507.
- [20] Altaf-Ul-Amin M et al. (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks, BMC Bioinformatics, 7, 207.