Mining Correlation Patterns of Taxa, Pathways and Environmental Factors with An Improved Weighted Network Community Detection Algorithm

Xiao-Ying Yan[1,2], Shao-Wu Zhang[1*], Ze-Gang Wei[1], Wei-Feng Guo[1]
[1] College of Automation, Key laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical Univerwsity, Xi'an 710071, China
[2] College of Computer Science, Xi'an Shiyou University, Xi'an 710065, China
*Corresponding Author, Email: 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. Generally, the functional capability and taxa structure are highly related with environment factors in microbial communities, which are hidden in these large amount sequences. However, most works used the canonical correlation analysis (CCA) method to research the correlative relationship among taxa, pathways and environmental factors. CCA is difficult to find which environmental factors are the major determinants of some special taxa and pathway. In this paper, we integrated 14 ocean metagenomes with geographical, meteorological and geophysicochemical data to construct the correlative weighted networks with Spearman correlation. By using an improved weighted network community detection algorithm, named as IWNCD, we find some special correlation patterns among taxa, pathways and environmental factors. Analysis of these patterns shows that the climatic factors such as temperature, sunlight, and correlated CO_2, and the nutrients such as chlorophyII and primary production are the main determining factors of the functional community composition; The growth and development of some special taxa are dependent on some main environmental factors such as sunlight, temperature, CO_2, primary production, dissolved oxygen, dissolved silicate; In addition, sampling sites more similar in geographic location have a greater tendency to be closer together based on their metabolic pathways.**

Keywords—marine microbe; taxa; pathway; weighted network; correlation pattern

I. INTRODUCTION

Microbial communities are the combinations of bacteria, archaea, fungi, yeasts, eukaryotes and viruses, often co-occurring in a single habitat, which carry out the majority of the biochemical activity on the earth. The metabolic process performed by microbes is important both for understanding and manipulating ecosystems. However specific ecological functional relationships among these microbial taxa and environment factors are largely unknown. This is partly due to the dilute, microscopic nature of the planktonic microbial community, which prevents direct observation of their interactions [1]. With the development of high-throughput DNA sequencing technologies to sample the genetic content of heterogeneous environments, a mass of reads of small-subunit rRNA gene (16S rRNA/18S rRNA) and DNA were generated. So, we can describe the compositions of microbial communities, their metabolic functional diversity and how communities change across space based on these sequence data [2]. 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[3-5], and often use the canonical correlation analysis (CCA) method to analyze the relationships among taxa, pathway and environmental factors[6-9]. CCA is difficult to find which environmental factors are the major determinants of some special taxa and pathway. In contrast, there has been far less attention focused on using sequence data to explore the direct or indirect relationship among microbial taxa and environments. Although some researchers used the network analysis to explore co-occurrence pattern in soil and ocean[1-2, 10-11], they just constructed the association networks to show the co-occurrence pattern, and did not further mine the networks to find the pattern structures.

Generally, the functional capability and taxonomic structure are highly related with environment factors in microbial communities, which are hidden in these large amount sequences. In this paper, we integrated 14 ocean metagenomes with geographical, meteorological and geophysicochemical data, and used Spearman method to compute the correlation among taxa, pathways and environmental factor for constructing the correlative weighted networks, and adopted an improved weighted network community detection algorithm to research the correlation patterns among taxa, pathways and environmental factors. The aim was to find which environmental factors are the major determinants of some special taxa and pathway.

II. MATERIAL AND METHODS

A. Dataset

Sequences and metadata (acidity, CO_2, chlorophyII, nitrate, dissolved oxygen, phosphate, primary production, salinity, dissolved silicate, sunlight, temperature, sample depth, water depth, mixed layer depth) from GOS (Global Ocean Sampling) Expedition [12-13] were downloaded from CAMERA[14].We

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

first filtered the data of the GOS Expedition to keep only those sites that used a filter size of 0.1-0.8μm (i.e. majority prokaryote samples). To ensure reliability of the following analysis, we selected 14 sites by filtering those with less than 5000 usable sequences, less than 33% pathway annotation rate of sequences, less than 53% taxon annotation rate of sequences and incomplete metadata. These sampling sites locate in North Atlantic, Mid-Atlantic and South Pacific.

B. Pathway and taxon annotation

There are two kinds of approaches to annotate the metagenomic sequences. One is the taxonomic annotation, many computational tools have been proposed for this task[15-16]. The other is functional annotation, which can be achieved by several tools [17-19]. For simplifying analysis, we obtained the taxonomic annotations at class level and functional annotations at pathway level (that is, KEGG Orthologous groups, KO) directly from MG-RAST web (http://metagenomics.anl.gov). Then, we can derive two abundance matrices: one is taxonomic abundance matrix with 68 classes * 14 samples, which contains 1,082,546 sequences, another is pathway abundance matrix with 134 KOs *14 samples, which contains 844,702 sequences.

C. Weighted correlative network modeling

In order to investigate the correlation among taxon and environmental factor, pathway and environmental factor, we used vector T_i, P_j, E_k to represent taxon, pathway and environmental factor, respectively.

$$T_i = [t_{1i}, t_{2i}, ..., t_{si}, \cdots \qquad (i = 1,2,...,68) \tag{1}$$

$$P_j = [p_{1j}, p_{2j}, ..., p_{sj}, \cdots \qquad (j = 1,2,...,134) \tag{2}$$

$$E_k = [e_{1k}, e_{2k}, ..., e_{sk}, \cdots \qquad (k = 1,2,...,14) \tag{3}$$

where t_{si} is the i-th taxon abundance value in the s-th sampling, that is, t_{si} equals the ratio of annotated sequence number N_{si} contained in i-th taxon and the total annotated sequence number N_s contained in the s-th sampling; p_{sj} is the j-th pathway abundance value in the s-th sampling, that is, p_{sj} equals the ratio of annotated sequence number N_{sj} contained in j-th pathway and the total annotated sequence number N_s contained in the s-th sampling. To reduce the bias of correlation analysis, the t_{si} and p_{sj} were set to zero if t_{si}<0.001 and p_{sj}<0.001. e_{sk} is the k-th environmental factor value in the s-th sampling. E_k represents the environmental factor variable such as CO_2 (E1),water depth (E2), chlorophyll (E3), acidity (E4), sample depth (E5), mixed layer depth based on temperature (E6), nitrate (E7), dissolved oxygen (E8), phosphate (E9), salinity (E10), dissolved silicate (E11), sunlight (E12), temperature (E13) and daily primary production (E14), which was normalized with zero-mean and standard deviation σ=1.

Spearman method was used to compute the pairwise correlations among taxon variables and environmental variables, pathway variables and environmental variables, respectively, and the permutation test was adopted to calculate the statistical significance. If P-value<0.01, the correlations among taxon and environmental variables, pathway and environmental variables are considered as strong correlation, that is, there is an edge to link the taxon (or pathway) variable with environmental variable. If the correlation coefficient value is used to represent the edge, we can construct two correlatively weighted networks: taxon-environmental factor network and pathway-environmental factor network.

D. Improved weighted network community detection algorithm

Lu et al. [20] borrowed the notion of conductance[21] defining the community *conductance function* and the *belonging degree*, and further proposed a community detection algorithm to mine the weighted networks. Their algorithm fits this kind of weighted networks in which the edge weight value is positive. However, some edge weight values in our taxon-environmental factor network and pathway-environmental factor network are negative. Then, we proposed an improved weighted network community detection algorithm, named as IWNCD.

Let G=(V, E) represents a weighted and undirected network, where V denotes the node set and E denotes the edge set. We first defined two functions: *belonging degree* and *conductance*.

For a community C and a node u, the *belonging degree* $B(u,C)$ of u belonging C is defined as:

$$B(u,C) = \sum_{v \in C} |w_{uv}| \Big/ k_u \tag{4}$$

where $k_u = \sum_{v \in N_u} |w_{uv}|$ is the weighted degree of node u, and N_u is the neighbor set of node u.

The *conductance* $\Phi(C)$ of community C is defined as:

$$\Phi(C) = E_C / (I_C + E_C) \tag{5}$$

where $I_C = \sum_{u,v \in C} |w_{uv}|$ is the sum of absolute weigh value of edges in C, $E_C = \sum_{u \in C, v \notin C} |w_{uv}|$ is the sum of absolute weigh value of edges on the boundary of C. With lower *conductance* $\Phi(C)$, more edge weights are within the community and the identified community is better.

The improved community detection algorithm for weighted networks can be described as follow. For a given weighted network G, we first choose those nodes connected by the highest weight edges as a candidate community C and calculate its *conductance* $\Phi(C)$. Then, the boundary adjacent node u of community C with the highest $B(u,C)$ is combined with C, and also compute $\Phi(C+u)$. If $\Phi(C+u) < \Phi(C)$, add node u to C forming a new community C'; otherwise, C is designated as a detected community. The whole process is repeated until the edge set is empty.

The whole process of matagenomic sequences annotation and detection of the correlation patterns of taxa-environmental factors and pathway-environmental factors was shown in the Figure 1.

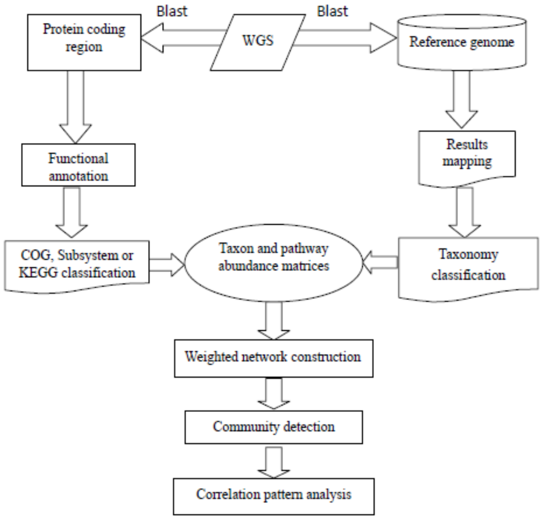
Blast
Blast
Protein coding region
WGS
Reference genome
Functional annotation
Results mapping
COG, Subsystem or KEGG classification
Taxon and pathway abundance matrices
Taxonomy classification
Weighted network construction
Community detection
Correlation pattern analysis

Fig.1. The flowchart of the taxonomy, pathway annotation and detection of correlation patterns of taxa-environmental factors and pathway-environmental factors

III. Results And Discussion

A. Variation in metabolic function corresponds to geographic regions

With metastats algorithm[22], we found 14 significantly different pathways among sampling sites. Using these 14 pathway abundance features to compute the pairwise correlation between sites, we observed significantly variation between the sampling sites as shown in Figure 2, where site pairs are color-coded according to their similarity. Additionally, based on the similarity of pathway abundance features, the 14 sampling sites were clustered into three groups: the North Atlantic, the Mid-Atlantic, and the Pacific, which is strong concordance with geographic location. We also analyzed the significant pathways, and found that the ko00280(Valine, leucine and isoleucine degradation), ko02020(Two-component system), ko00330 (Arginine and proline metabolism), ko00630(Glyoxylate and dicarboxylate metabolism, ko00350(Tyrosine metabolism), ko00140(Steroid hormone biosynthesis), ko04113(Meiosis-yeast), ko00930(Caprolactam degradation), ko00983(Drug metabolism-other enzymes) are higher enriched in North Atlantic samples, and most of these pathways are involved in Amino acid metabolism, Signal transduction, Carbohydrate metabolism, Cell growth, Xenobiotics biodegradation and metabolis; the ko0970(Aminoacyl-tRNA biosynthesis), ko0130(Ubiquinone and other terpenoid-quinone biosynthesis), ko4066(HIF-1 signaling pathway), ko0040(Pentose and glucuronateinterconversions), ko0906(Carotenoid biosynthesis), ko0750(Vitamin B6 metabolism), ko3060(Protein export), ko0440(Phosphonate and phosphinate metabolism) are highly enriched in Pacific samples, and most of these pathways are involved in Translation, Metabolism of cofactors and vitamins, Signal transduction. These results suggest that sampling sites more similar in geographic location have a greater tendency to be closer together based on their metabolic pathways.

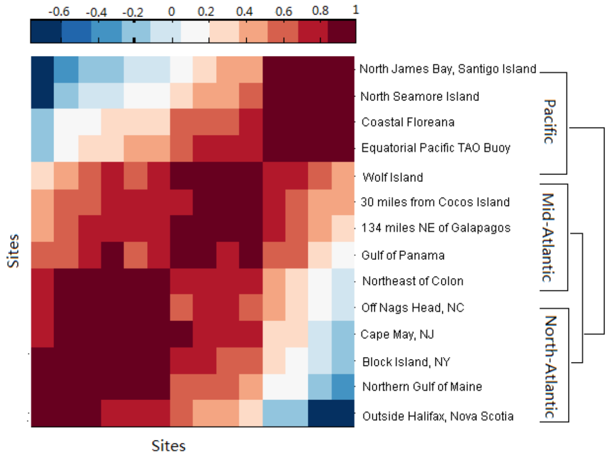

Fig.2. Site-site correlations and clustering results with significantly different pathway abundance features

B. The pathway-environment correlation patterns detected by IWNCD

Pervious researches have shown a clear impact of environmental conditions on the functional composition of microbial community [8, 23]. Here, we used the IWNCD algorithm to investigate which environmental factors are the main drivers in the pathways. The correlation patterns of pathways and environmental factors detected by IWNCD were shown in Figure 3. Four patterns (or communities) in Figure 3 show that some environmental factors are strongly correlative with some special pathways. For example, in C1 community, the environmental factors dissolved oxygen (E8) and dissolved silicate (E11) are strongly correlative with ko00040, ko00330, ko00930, ko00960, ko00983, ko04080, ko04113, ko04144, ko03040, ko05204. In C2 community, the environmental factors CO2(E1), chlorophyII(E3), sunlight(E12), temperature(E13) and primary production(E14) are strongly correlative with ko00130, ko00140, ko00195, ko00196, ko00240, ko00280, ko00350, ko00362, ko00450, ko00630, ko00710, ko00860, ko00900, ko00906, ko00930, ko00970, ko02020, ko02060, ko03030, ko03420, ko04110, ko04115, ko04151, ko05200 and ko05340. We also counted the number of pathways whose correlative coefficient absolute value with some special environmental factors are more than 0.6 ($|r| > 0.6$), and found that there are 42, 26, 24, 19, 13, 11, 8 and 6 pathways correlating with temperature (E13), chorophyII (E3), CO2 (E1), primary production (E14), silicate (E11), dissolved oxygen (E8), sunlight (E12) and water depth (E2), respectively. These results and C1, C2, C3 communities in Figure 3 suggest that the climatic factors such as temperature and sunlight (and correlated CO_2), and the nutrients such as chlorophyII and primary production are the main determining factors of the functional community composition.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

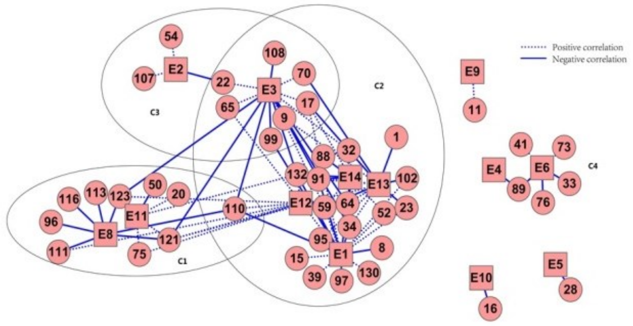

Fig. 3. The correlation patterns of pathway and environmental factors detected by IWNCD

C. *The taxa-environment correlation patterns detected by IWNCD*

The correlation patterns of taxa and environmental factors detected by IWNCD were shown in Figure 4. Five patterns (or communities) in Figure 4 show that some environmental factors are strongly correlative with some special taxa. For example, in C1 community, the primary production (E14) has strongly negative correlation with Aquificae, Bacilli, Chlorobia, Chloroflexi, Clostridia, Deferribacteres, Deinococci, Deltaproteobacteria, Dictyoglomia, Fusobacteriia, Methanobacteria, Methanococci, Negativicutes, Thermococci, Thermoprotei, Thermotogae, which indicates that these marine microbes draw the nutrients from the primary production for growing. In community C5, the climatic factors sunlight (E12), temperature (E13) and correlated CO2 (E1) are strongly correlative with Elusimicrobia, Erysipelotrichi, Halobacteria, Mollicutes, Proteobacteria, Spartobacteria, Synergistia and Verrucomicrobiae.

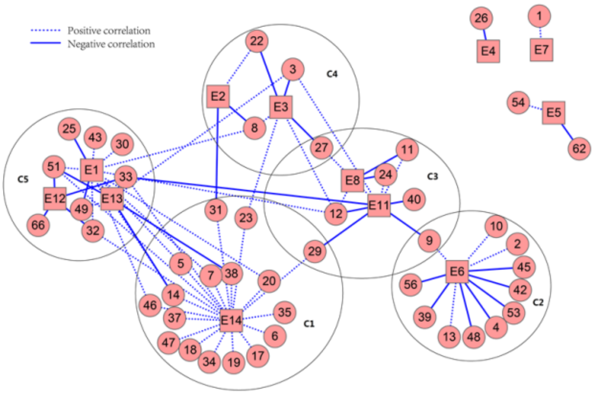

Fig. 4. The correlation patterns of taxa and environmental factors detected by IWNCD

IV. Conclusions

Mining the marine taxa and functional correlation patterns and diversity is a key step for exploiting the marine resources. Considering that the environmental factors strongly effect the marine microbes, we integrated different metagenomic data from 14 global ocean sampling sites, and used an improved weighted network community detection algorithm (IWNCD) to research which environmental factors are the major determinants of some special taxa and metabolic pathway. The results show that the climatic factors such as temperature, sunlight, and correlated CO_2, and the nutrients such as primary production and chlorophyII are the main determining factors of the functional community composition; The growth and development of some special taxa are dependent on the main environmental factors such as sunlight, temperature, CO_2, primary production, dissolved oxygen, dissolved silicate; In addition, sampling sites more similar in geographic location have a greater tendency to be closer together based on their metabolic pathways.

Acknowledgment

This paper was supported by the National Natural Science Foundation of China (61170134) and the graduate starting seed fund of Northwestern Polytechnical University (Z2014145, Z2014152).

References

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

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

[3] 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.

[4] 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.

[5] D.L. Kirchman, M.T. Cottrelland, 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.

[6] P.V. Patel, T.A. Gianoulis, R.D. Bjornson, *et al*., “Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families". Genome Research, vol.20, pp. 960-971,2010.

[7] T.A. Gianoulis, J. Raes, P.V. Patel, *et al*. "Quantifying environmental adaptation of metabolic pathways in metagenomics". Proc. Natl. Acad. Sci. USA, vol. 106, pp. 1374-1379, 2009.

[8] J.Raes, L.Letunic, T.Yamada, *et al*., “Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data,” Molecular Systems Biology, vol.7, pp. e473, 2011.

[9] X.Tian, F.Gong, S. Zhang, “Detect taxonomy-specific pathway associations with environmental factors using metagenomic data,” The 7th international conference on Systems Biology (ISB), HuangShan, China, August 23-25, 2013.

[10] J. Zhou, Y. Deng, F. Luo, *et al*.,“Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO_2,” mBio, vol.2, pp. e0122-11, 2011.

[11] J.A. Gilbert, J.A. Steele, J.G. Caporaso, *et al*.,“Defining seasonal marine microbial community dynamics,” The ISME Journal, vol. 6, pp. 298-308, 2012.

[12] D.B. Rusch, A.L.Halpern, G. Sutton, *et al*. “The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific”. PLOS Biology, vol. 5, pp. e77, 2007.

[13] H.Garcia, R.Locarnini, T. Boyer, J. Antonov, World Ocean Atlas 2005. In NOAA Atlas NESDIS 63 (ed. S Levius), pp. 342. U.S. Government Printing Office, Washington, DC, 2006.

[14] R. Seshadri, S.A. Kravitz, L. Smarr, et al., “CAMERA: A community resource for metagenomics,” PLOS Biology, vol.5, pp.e75, 2007.

[15] H. Jiang, L. An, S.M. Lin, *et al*. "A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads". PlOS ONE, vol. 7, pp. e46450, 2012.

[16] B.Yang, Y.Peng, H.Leung, *et al*. "MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation". Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, pp. 170-179, Niagara Falls, NY, USA, August 2-4, 2010

[17] F. Gori, G. Folino, M.S.M. Jetten, *et al*., "MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks". Bioinformatics, vol. 27, pp. 196-203, 2011.

[18] F. Meyer, D.Paarmann, M. D'souza, *et al*. "The metagenomics RAST server a public resource for the automatic phylogenetic and functional analysis of metagenomes". BMC Bioinformatics, vol. 9, pp.386, 2008.

[19] S. Mitra, P. Rupek, D.C. Richter, *et al*., "Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG". BMC Bioinformatics, vol. 12 , pp. S21, 2011.

[20] Z. Lu, Y. W, G. Cao. "Community detection in weighted networks: algorithms and application," 2013 IEEE Interantional Conference on Pervasive Computing and Communications(PerCom), San Diego, March 18-22, 2013.

[21] J. Leskovec, K. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," International World Wide Web Conference Committee (IW3C2), Raleigh, North Carolina, USA, April 26–30, 2010.

[22] J.R.White, N. Nagarajan, M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples". PLOS Computational Biology, vol. 5, pp. e1000352, 2009.

[23] E.A. Dinsdale, R.A. Edwards, D. Hall, et al. "Functional metagenomic profiling of nine bioes". Nature, vol. 452, pp. 629-632, 2008.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE