Detect taxonomy-specific pathway associations with environmental factors using metagenomic data

Xue Tian, Fuzhou Gong* and Shihua Zhang*
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
*Corresponding authors. Email: zsh@amss.ac.cn or fzgong@amt.ac.cn

***Abstract*—In microbial communities, the taxonomic structure and functional capability are highly related. We proposed a method by considering the combination of taxa and functional categories to explore the ecological mechanisms of microbial communities. Using GOS metagenomic samples, we tested this idea and its effectiveness. The combination of taxonomies and functional groups could reflect the difference between habitats and may help to explain the combination adaptability of microbes to environment.**

I. Introduction

Microbes comprise the most of organisms, living everywhere throughout the ocean, the soil, and human bodies and so on. They play important roles in the recycle of nutrients, the degradation of toxins [1] and the maintenance of human health [2]. However, the organization and function of microbial community remain mysterious. The diversity of microbes and their interactions with environmental factors are not well known due to poor cultivability and their complexity [3].

Fortunately, recent advances in sequencing technologies [4] have allowed us to investigate the microbes that inhabit oceans, human bodies and elsewhere. Recent metagenomic studies began to study the composition of microbial communities and have found close relationship between phylogenetic diversity and environmental factors such as temperature and latitude [5,6]. Several studies focused on the interactions between environmental factors and metabolic functions [7]. They found that many of the environment dependent pathways were associated with energy conservations such as photosynthesis, oxidative phosphorylation, carbon and nitrogen fixation.

We note that most studies investigated the taxonomic and functional diversity individually. However, the community structure and metabolic functions are highly dependent. Different species tend to play different roles and co-exist in a specific environment [8]. So, when identifying the relationship between microbes and habitats, the combination adaptability of taxonomic structures and metabolic capabilities to environmental changes should be considered. We proposed a method to identify metagenome-environment associations considering the effect of species abundance and functional enrichment simultaneously. Many previous methods calculate only species (or pathways) abundance and derive a one dimension vector for each metagenomic sample (we call this a univariate method). Different from this type of methods, we treat the group of taxa and pathways as a binary variable and get an abundance matrix for each sample, where each row represents a pathway and each column stands for a taxa. Besides, by correlation analysis, the two-tuples of taxonomy and pathway can be revealed to explain the ecological mechanism of microbial communities.

To test the effectiveness of our proposal, we applied our method on an ocean metagenomic dataset. The ocean comprises 71% of the earth's surface and is the largest ecosystem on earth. Vast numbers of bacteria and plankton occur both at the surface and in deep ocean waters. The global ocean sampling project expedition [9] provide us a comprehensive dataset to shed light on the role of marine microbes. Here, we used 59 GOS samples from MGRAST to explain the procedure and effectiveness of our method. The DNA sequences of whole community were annotated with both taxonomies and pathways. We obtained a taxonomy and pathway abundance matrix. This matrix was used to study the association between environment and taxonomy specific pathways. It showed that taxonomy*function groups can reflect the difference of sampling habitats. Taxonomy specific pathways with differential abundance between coastal and open ocean sites, or significant correlations with temperature are also identified.

II. Materials and Methods

A. Materials

We utilized the Global Ocean Sampling (GOS) expedition from MG-RAST (http://metagenomics.anl.gov/). GOS project is the largest published one for metagenomics on marine environment. It gathers ocean surface samples across a transect from the North Atlantic through the Panama Canal and ending in the South Pacific [9]. This project generated approximately eight billion nucleotides present in more than seven million DNA fragments, providing an unprecedented resource to study marine microbes and their interactions with the environment. Currently, 88 samples can be available from MG-RAST. We obtained 59 samples by filtering those with less 5000 usable sequences to ensure reliability for the following analysis.

B. Methods

Many studies have explored the relationship between taxa or functional categories and environment factors. Here, we propose a method to identify metagenome-environment associations by considering the combinatorial effect of taxonomy abundance and functional enrichment (Figure 1). Taxonomy-specific functional categories will be identified to explain the ecological mechanism of microbial communities.

We processed metagenomic sequences of the whole community started with two types of annotations. One is the

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

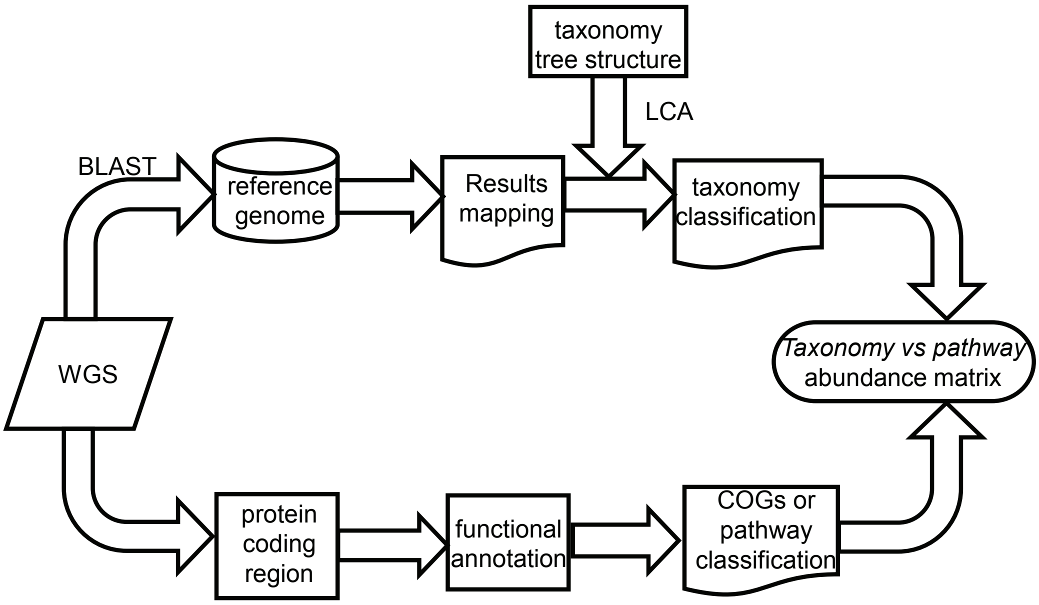

Fig. 1. The flowchart on the generation of taxonomy and pathway abundance matrix for a sample. Metagenomic sequencing of whole community DNA was firstly processed in two ways. 1) Taxonomic annotation using LCA based methods or Metacluster, MTR and so on. 2) Functional category classification. Protein coding regions of the reads are predicted, annotated and assigned to their functional categories, for example, COGs or KEGG pathways. 3) The last step is to integrate the taxonomic and functional annotation and obtain a taxonomy and pathway abundance matrix for each sample. The marginal distributions are actually the taxonomic composition or functional profile considered by previous methods.

taxonomic annotation, many computational tools have been proposed for this task [10-12]. The other is functional annotation, i.e, mapping protein coding regions to orthologous groups, which can be achieved by several tools [13,14]. To simplify our analysis, we obtained the taxonomic annotations at class level and functional orthologous groups (or KEGG Orthologous groups, KO for short) directly from MG-RAST. Finally, for each sequence, we combine its taxonomic and functional annotations together and get taxonomy and pathway abundance matrix for each sample. Taxonomy and pathway abundance for a sample was calculated by summing the number of assignments that belongs to the taxonomy and pathway and standardized by the total number of assignments for this sample. In our study, we derive a matrix with 166 classes versus 7660 KOs for each sample. The marginal distribution of this matrix corresponds to the phylogenetic or functional abundance as considered in previous studies [6,7].

This taxonomy and pathway abundance matrix can be specified at different levels, using the hierarchy information of taxonomic trees from NCBI and functional linkage definition (e.g., 153 pathways) from KEGG. The abundance of a higher order node (e.g., class*pathway) is calculated by summing the abundance of all branch members (e.g., class*KO) in the hierarchy structure. The abundance of nodes that are members of more than one higher level nodes (e.g., some enzymes can take part in multiple pathways) are equally split among higher nodes. We also get its taxonomic and functional abundance matrix at class*pathway level and derived a 158 classes*148 pathways matrix for each sample. In a few places, we combine the results of these two levels together and we analyze at the class*pathway level in most cases.

C. Wilcoxon-Mann-Witney test to detect differential abundance groups

We use non-parametric Wilcoxon-Mann-Witney test to detect taxa or taxonomy specific pathways between 18 coastal and 20 open ocean habitats with significant differential abundance. For taxa, we only test those with average abundance >0.1% among 38 habitats (p-value<0.05). For taxonomy specific pathways, p-values are adjusted to produce a Benjamini false discovery rate.

D. Correlation analysis

Spearman correlation test was used to identify the correlation between taxonomy specific pathways and temperature. For those groups, whose non-zero rate (or occurrence rate) is less than 25% (e.g. 15 in 57 samples) was removed from correlation analysis. Obtained p-values for these tests were adjusted for multiple testing using Benjamini-Hochberg false discovery rates [13]. Taxonomy specific pathways are further filtered according to adjusted p-values.

III. Results

We began to analyze the difference of taxonomic and functional structure in all samples and try to relate these differences to available environmental factors.

A. Taxonomy-specific functional profiles are effective to reflect sampling differences

Principal components analysis (PCA) has been comprehensively used for analyzing the differences between metagenomic samples [7]. To assess the performance of taxonomic and functional composition as community descriptor, PCA is used for analyzing the diversity and consistency of these 59 samples

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

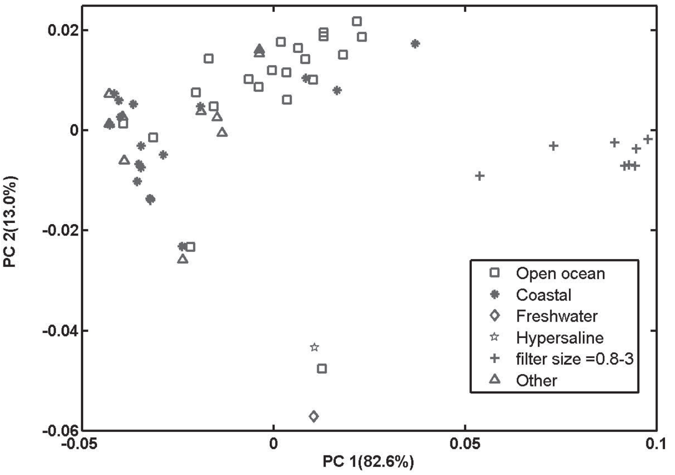

Fig. 2. Visualization of sample distributions using the two leading PCA dimensions. The variations explained by the first two components are indicated on the two axes respectively.

TABLE I.

Taxonomy abundance for GOS samples. 10 dominant taxa are listed with average abundance in coastal, open ocean and entire GOS samples. The italic ones are those differentially enriched between coastal and open ocean habitats with p-value <0.05, by Wilcoxon-Mann-Witney test.

Taxonomy	Coastal	Open Ocean	59 samples
Alphaproteobacteria	*57.1%*	52.2%	47.6%
Cyanobacteria	8.7%	*21.6%*	19.9%
Gammaproteobacteria	*10.7%*	7.7%	9.8%
Flavobacteria	*6.4%*	4.5%	5.1%
Actinobacteria	2.9%	2.1%	3.6%
Betaproteobacteria	*2.2%*	1.3%	2.4%
unclassified Bacteria	*1.8%*	1.6%	1.5%
unclassified Viruses	*0.9%*	0.7%	0.9%
Deltaproteobacteria	0.5%	0.7%	0.7%
Clostridia	0.7%	0.7%	0.6%

revealed by the combinatorial characteristics. We obtained 50 class*pathway features which have the largest average abundance among 59 samples. As showed in Figure 2, the samples can generally be classified into different groups. Eight plus signs indicate samples collected within size fraction, 0.8-3 μm. It was pointed that more eukaryotic organisms are included in this range, which effect the accuracy of results and are often excluded by previous studies [7,15]. Three sites are obviously deviated from others, among which two are the hypersaline sample GS033 (salinity 6.54) and freshwater sample GS020 (salinity 0.01). This may further confirm the theory that phylogenetic composition is mainly determined by salinity in extreme environments [16]. Most of the remainders are involved in coastal and open ocean sites, also some harbor, reef, and mangrove samples. Generally, we can see that the costal samples are separated from open ocean samples. This demonstrates that taxonomy specific pathway abundance is informative to discriminate samples from different habitats and collected with different filter size, and provides a new perspective to further explore metagenomes.

B. Identify differentially enriched taxonomy specific pathways between coastal and open ocean habitats

As we have mentioned that the marginal distribution derived from the taxonomy and pathway abundance matrix is actually the taxonomic profile of the community. We obtained the taxon distribution for 59 samples including 18 coastal habitats and 20 open ocean habitats in GOS (Table 1). The average abundance of taxa was shown by class in Figure 2. We get similar inference of marine microbial diversity with other studies (e.g., that by using 16S RNA gene sequences) to explore prokaryotic biodiversity in surface marine waters [17-19]. Alphaproteobacteria dominants the marine community (47.6% of total classified sequences) and Cyanobacteria is the second sdominant taxon (19.9%). Cyanobacteria is more abundant than reported in previous studies. This may due to the effect of genome size [20]. We should note that Viruses and Archaea only make up 0.9%, 0.3% of the total diversity of GOS respectively. Some taxa have differential abundance between coastal and open ocean habitats based on Wilcoxon-Mann-Witney test ($p < 0.05$) (Table 1). Flavobacteria, Alphaproteobacteria, Viruses are more abundant in coastal than in open ocean communities, while Cyanobacteria was more prevalent in open ocean habitats, as in previous study [18]. Besides, we also observe that Gammaproteobacteria, Betaroteobacteria and viruses have differential abundance in coastal sites.

Using our method, we can also detect taxonomy specific pathways that have differential abundance between coastal and open ocean habitats (see Methods). These findings probably can help to explain which pathways contribute more to taxonomic divergences. We observe top 100 taxonomy specific pathways (q-value <0.0016) in Figure 2, that are differentially enriched in one habitat. Almost all these taxonomy specific pathway (95 in 100) belongs to the five differential abundant taxa. Most of these pathways belong to Cyanobacteria. This may because of its prevalent and important role in marine ecosystems [21]. Cyanobacteria are fundamental for oceanic primary production [19] and carbon and nitrogen fixation [22]. We find that ko00860 (Porphyrin and chlorophyll metabolism), ko00195 (Photosynthesis), ko00720 (Carbon fixation pathways), ko00190 (Oxidative phosphorylation) are all differentially abundant in open ocean habitats. Besides, many pathways involved with carbohydrate metabolism (ko00010, ko00020, ko00030, ko00040, ko000051 e.g.) and amino acid metabolism (ko00340, ko00260, ko00330, ko00400, ko00250 e.g.) are also identified. All these pathways are fundamental and important for microbes. For Alphaproteobacteria, we detected 15 pathways more abundant in coastal habitats. Three are related to carbohydrate metabolism (ko00630, ko00640, ko00650) and three for amino acid metabolism (ko00260, ko00290, ko00350). Two pathways for xenobiotics biodegradation are enriched in coastal habitats. This may help Alphaproteobacteria to resist toxin more abundant in coastal sites. For viruses, we identify that three pathways are more prevalent in coastal communities. They include ko00480 for amino acid metabolism and ko00230 for nucleotide metabolism. These pathways are crucial for the structure and function of viruses [23].

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

TABLE II.

Taxonomy specific pathways differentially abundant between coastal and open ocean habitat. Taxonomy specific pathways differentially abundant are ranked by corrected p-value. Top 100 are remained. Rows are taxonomies. Columns are pathways. Column two is total identified pathways for the taxonomy. Column three to six are pathways falling into the corresponding functional categories.

Taxonomy-pathway	Total	Energy metabolism	Carbohydrate metabolism	Amino acid metabolism	Nucleotide metabolism
Cyanobacteria	51	6(ko00710, ko00190, ...)	10(ko00010, ko00020, ...)	11(ko00340, ko00260, ...)	2(ko00230, ko00240)
alphaproteobacteria	15	2(ko00680, ko00720)	3(ko00630, ko00640, ko00650)	3(ko00260, ko00290, ko00350)	0
Flavobacteria	14	1(ko00190)	3(ko00010, ko00500, ko00650)	3(ko00340, ko00290, ko00250)	1(ko00240)
gammaproteobacteria	12	1(ko00190)	2 ko00010, ko00620)	3(ko00410, ko00290, ko00250)	0
viruses	3	0	0	1(ko00480)	1(ko00230)

C. *Environmental factors can have different effects on taxonomy specific pathways*

57 GOS samples with available temperature information are used to analyze the correlation between taxonomy specific pathways and temperature (both are at class*pathway level and class*KO level). We identify 355 taxonomy specific pathways which are significantly related with temperature (with corrected p-value<0.1) (Figure 2) including 65 pathways for Cyanobacteria, 60 for Gammaproteobacteria and 44 pathways for Alphaproteobacteria. We see that fewer ones belong to Alphaproteobacteria in spite of its prevalence in marine habitat. For most taxonomies, only few pathways are detected which may result from the diversity of their abundance. For those significant correlations, both positive and negative associations are detected. We can see that environmental factors can have different effects on pathways among different taxonomies and even within the same taxonomy. Our further analysis focuses on Alphaproteobacteria and Cyanobacteria because of their dominance in marine habitat. For both taxa, top 40 pathways filtered by corrected p-values are summed up by their functional categories in Table 3. Most pathways belong to amino acid Metabolism, carbohydrate metabolism and energy metabolism. All are fundamental pathways for microbes as detected in [7]. For Cyanobacteria, we find that Porphyrin and chlorophyll metabolism (ko00860, spearman $r = 0.50$) and Photosynthesis (ko00195, $r = 0.39$) are positive correlated with temperature. Chlorophyll is a component of the photosynthetic machinery; it absorbs light energy and is involved in energy transfer in the course of photosynthesis 1 and 2 [24]. Besides, pathways involved with Carbon fixation (ko00720, $r = 0.47$), Nitrogen metabolism (ko00910, $r = 0.47$), Oxidative phosphorylation (ko00190, $r = 0.46$) are also positive related with temperature.

We do correlation analysis from class and KO levels and detect cyanobacteria specific enzymes which are significantly correlated with temperature (see Table 4). This may help to explain at more specific levels and find enzymes that contribute more to the correlation. For example, hydroxymethylbilane synthase (K01749) taking part in chlorophyll [25] is found to be positively correlated with temperature. Phosphoglycerate kinase (PGK, K00927) and fructose-bisphosphate aldolase (FBA, K01623) relating to glycolytic and photosynthetic reactions in photosynthetic organisms are also identified. Both enzyms have been found in Synechocystis sp.PCC6803 [26,27]. The activity of PGK has already been shown to be temperature-sensitive [28]. Cyanobacterial FBA expressed in transgenic tobacco plants can enhance photosynthetic efficiency and growth characteristics. Besides some enzymes contributing for amino acid metabolism have also been detected.

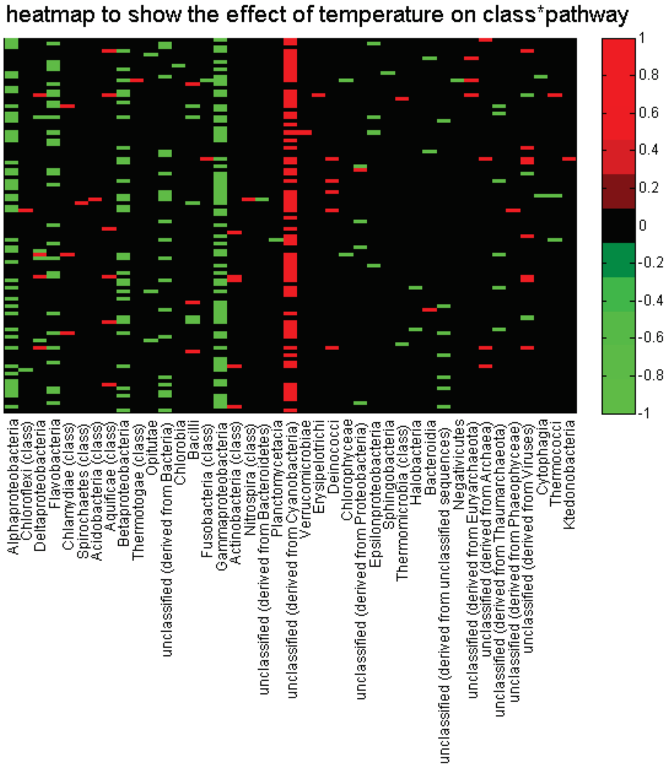

Fig. 3. Heatmap for showing the effect of temperature on taxonomy specific pathways. Only 355 class*pathway groups (corrected p-value < 0.1) are colored by the signs of their correlation coefficient, red for 1 and green for -1. Others are all black for 0.

IV. Discussion

In microbial communities, the taxonomic structure and functional capabilities are highly related. Instead of using taxonomies or pathways individually, we proposed a method by considering the combination of taxa and functional categories to explore the ecological mechanisms of microbial communities. Using GOS metagenomic samples, we tested the workflow and effectiveness of this method. The combination of taxonomies and functional groups could reflect the difference between habitats and may help to explain the combination adaptability of microbes to environment.

There are still some problems with this method. We note

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

TABLE III.

The distribution of top 40 pathways for cyanobacteria and alphaproteobacteria, which are significantly correlated with temperature and filtered by corrected p-values.

taxonomy-pathway	Carbohydrate metabolism	Amino-acid metabolism	Energy metabolism	Nucleotide metabolism	Cofactor vitamins
Cyanobacteria	10	13	6	3	3
alpha-proteobacteria	8	10	6	0	5

TABLE IV.

The pathways and enzymes for cyanobacteria correlated with temperature.

Pathway	Rank	Functional annotations	Enzymes (corrected p <0.1)
ko00400	1	Phenylalanine, tyrosine and tryptophan biosynthesis	K01626 K00210 K01609
ko00860	2	Porphyrin and chlorophyll metabolism	K00798 K01749
ko00230	4	Purine metabolism	K03060 K01756 K00759 K00524 K00860
ko00051	5	Fructose and mannose metabolism	K00847 K01623 K00850
ko00010	7	Glycolysis or Gluconeogenesis	K00162 K00927 K01623 K00850
ko00910	8	Nitrogen metabolism	K02274 K01455 K01092
ko00710	14	Carbon fixation in photosynthetic organisms	K00927 K01623
ko00190	16	Oxidative phosphorylation	K02108 K05585 K02274 K05572 K05575

that different taxonomies may have different metabolic pathway composition. However, we download the pathway list directly from KEGG and used it for all taxonomies. This may result in artifacts in pathway analysis and needs to be further considered in similar analysis. Furthermore, when calculating the taxonomy and pathway abundance by read counts, there may exit biases due to different sizes of genomes and pathways. There are also overlaps among pathways; but we simply assigned KOs equally to all possible pathways. With the rapid accumulation of metagenomic data, we believe that these biases could be improved. We expected the proposed strategy will be more promising in the near future.

Acknowledgement

This project was supported by the Special Foundation of President of AMSS at CAS for ‘Chen Jing-Run’ Future Star Program and the Foundation for Members of Youth Innovation Promotion Association, CAS and the Scientific Research Foundation for ROCS, SEM.

References

[1] Arrigo KR. *Marine microorganisms and global nutrient cycles*. Nature, 2004, 437(7057): 349-355.

[2] Li M, Wang B, Zhang M, *et al. Symbiotic gut microbes modulate human metabolic phenotypes*. Proceedings of the National Academy of Sciences, 2008, 105(6): 2117-2122.

[3] O’Connell D. *A global unculture*. Nature Reviews Microbiology, 2006, 4(6): 418-419.

[4] Shendure J, Ji H. *Next-generation DNA sequencing*. Nature biotechnology, 2008, 26(10): 1135-1145.

[5] Fuhrman J A, McCallum K, Davis A. *Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans*. Applied and Environmental Microbiology, 1993, 59(5): 1294-1302.

[6] Fuhrman JA, Steele JA, Hewson I, *et al. A latitudinal diversity gradient in planktonic marine bacteria*. Proceedings of the National Academy of Sciences, 2008, 105(22): 7774-7778.

[7] Gianoulis TA, Raes J, Patel PV, *et al. Quantifying environmental adaptation of metabolic pathways in metagenomics*. Proceedings of the National Academy of Sciences, 2009, 106(5): 1374-1379.

[8] Tyson GW, Chapman J, Hugenholtz P, *et al. Community structure and metabolism through reconstruction of microbial genomes from the environment*. Nature, 2004, 428(6978): 37-43.

[9] Rusch DB, Halpern AL, Sutton G, *et al. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific*. PLoS biology, 2007, 5(3): e77.

[10] Jiang H, An L, Lin SM, *et al. A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads*. PloS one, 2012, 7(10): e46450.

[11] Yang B, Peng Y, Leung H, *et al. MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation*. Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. ACM, 2010: 170-179.

[12] Gori F, Folino G, Jetten MSM, *et al. MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks*. Bioinformatics, 2011, 27(2): 196-203.

[13] Meyer F, Paarmann D, D’souza M, *et al. The metagenomics RAST serverCa public resource for the automatic phylogenetic and functional analysis of metagenomes*. BMC bioinformatics, 2008, 9(1): 386.

[14] Li W. *Analysis and comparison of very large metagenomes with fast clustering and functional annotation*. BMC bioinformatics, 2009, 10(1): 359.

[15] Patel PV, Gianoulis TA, Bjornson RD, *et al. Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families*. Genome research, 2010, 20(7): 960-971.

[16] Lozupone CA, Knight R.*Global patterns in bacterial diversity*. Proceedings of the National Academy of Sciences, 2007, 104(27): 11436-11440.

[17] Pommier T, Canback B, Riemann L, *et al. Global patterns of diversity and community structure in marine bacterioplankton*. Molecular ecology, 2007, 16(4): 867-880.

[18] Biers EJ, Sun S, Howard EC. *Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome*. Applied and environmental microbiology, 2009, 75(7): 2221-2229.

[19] Giovannoni S, Rappe MS. *Evolution, diversity, and molecular ecology of marine prokaryotes*. Wiley Series in Ecological and Applied Microbiology, 2000.

[20] Konstantinidis KT, Tiedje JM. *Trends between gene content and genome size in prokaryotic species with larger genomes*. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 3160-3165.

[21] Hess WR. *Cyanobacterial genomics for ecology and biotechnology*. Current opinion in microbiology, 2011, 14(5): 608-614.

[22] Zehr JP, Waterbury JB, Turner PJ, *et al. Unicellular cyanobacteria fix*

N2 in the subtropical North Pacific Ocean. Nature, 2001, 412(6847): 635-638.

[23] Caspar DLD, Klug A. *Physical principles in the construction of regular viruses. Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press, 1962, 27: 1-24.

[24] Reinbothe C, Bakkouri ME, Buhr F, *et al. Chlorophyll biosynthesis: spotlight on protochlorophyllide reduction*. Trends in plant science, 2010, 15(11): 614-624.

[25] Silva PJ, Ramos MJ. *Comparative density functional study of models for the reaction mechanism of uroporphyrinogen III synthase*. The Journal of Physical Chemistry B, 2008, 112(10): 3144-3148.

[26] Tsukamoto Y, Fukushima Y, Hara S, *et al. Redox control of the activity of phosphoglycerate kinase in Synechocystis sp.*. Plant and Cell Physiology, 2013.

[27] Nakahara K, Yamamoto H, Miyake C, *et al. Purification and characterization of class-I and class-II fructose-1, 6-bisphosphate aldolases from the cyanobacterium Synechocystis sp. PCC 6803*. Plant and Cell Physiology, 2003, 44(3): 326-333.

[28] Pal B, Pybus B, Muccio DD, *et al. Biochemical characterization and crystallization of recombinant 3-phosphoglycerate kinase of Plasmodium falciparum. Biochimica et Biophysica Acta (BBA)*. Proteins and Proteomics, 2004, 1699(1): 277-280.

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