Manifold learning reveals nonlinear structure in metagenomic profiles

Xingpeng Jiang*, Xiaohua Hu*, Huiyu Shen†, Tingting He†
*College of Information Science and Technology, Drexel University,
Philadelphia, PA, United States
xpjiang@drexel.edu xh29@drexel.edu
† Central China Normal University
Wuhan, Hubei, P.R.China
shenhy8@hotmail.com tthe@mail.ccnu.edu.cn

***Abstract*—Using metagenomics to detect the global structure of microbial community remains a significant challenge. The structure of a microbial community and its functions are complicated not only because of the complex interactions among microbes but also their complicate interacting with confounding environmental factors. Recently dimension reduction methods such as Principle component analysis, Non-negative matrix factorization and Canonical correlation analysis have been employed extensively to investigate the complex structure embedded in metagenomic profiles which summarize the abundance of functional or taxonomic categorizations in metagenomic studies. However, metagenomic profiles are not necessary to meet the "Assumption of Linearity" behind these methods. Therefore it is worth to investigate how nonlinear methods can be utilized in metagenomic studies. In this paper, a nonlinear manifold learning method– Isomap is used to visualize and analyze large-scale metagenomic profiles. Isomap was applied on a large-scale Pfam profile which are derived from 45 metagenomes in Global Ocean Sampling expedition. In our result, a novel nonlinear structure of protein families is identified and the relationships among the identified nonlinear components and environmental factors of global ocean are explored. The results indicate the strength of nonlinear methods in learning the complex microbial structure. With the coming of the huge number of new sequenced metagenomes, nonlinear methods like Isomap could be necessary complementary tools to current widely used methods.**

***Keywords*-Nonlinear dimension reduction; metagenomic profile; Isomap; non-negative matrix factorization; principle component analysis**

I. Introduction

Understanding the structure of microbial communities is the key to elucidate the function of microbial ecology which is relevant to environmental conservation and human health. However, most of the species in a given microbial community are unculturable thus their functions and the interactions among them are very hard to investigate. In the last decade, metagenomics that sequence the genomic material from complex assemblages of microbial communities of bacteria, archaea, viruses et al. has yielded new insights into the structure of microbial ecology [1], [2]. Metagenomics is becoming a powerful tool to investigate the fundamental mechanisms of global ecology [3]–[7] and human diseases [8]–[10].

The accumulation of large-scale metagenomes dataset bring about the significant challenge for data analysis and visualization [11]. A metagenomic profile which summarized the abundance of functional or taxonomic categorizations derived from metagenome sequences could have thousands species or functional groups. Therefore, reducing the dimension of a metagenome dataset helps to visualize the organization of samples or functional and taxonomic groups, thus facilitates the biological interpretation of acquired data. Principle component analysis (PCA) [12] has been used in reducing the dimension of metagenomic profiles which summarized the abundance of functional or taxonomic categorizations derived from metagenome sequences [13]. For example, it was used recently in visualizing the enterotypes of human gut microbiome [14]. There is also a method called Principle coordinate analysis (PCoA–also called Multidimensional scaling, MDS) [15] which is used to visualize dissimilarities of data instead of the similarity in PCA. PCoA has been used for visualizing the relationships among the large number of taxonomic or functional groups. The widely used UniFrac has adopted PCoA as a standard visualization technology for exploring taxonomic relationships in microbial communities [16]. Recently we proposed a dimension reduction framework based on non-negative matrix factorization (NMF) [17] to gain a different and complementary perspective on relationships between functions, environment, and biogeography [18], [19]. To investigate the correlations among environmental variables and functions in global ocean metagenomes, canonical correlation analysis (CCA) has been proposed for visualize the linear relationships of environmental factors and functional categorizations [13], [20].

These methods are usually based on the assumption that the relationships embedded in a metagenomic dataset are linear in a Euclidean space. Under these frameworks, metagenome samples can be approximated ed as a linear combination of several reduced components (mixed sign or non-negative) or mapped to a low-dimension space by a linear transformation. However, this "Assumption of Linearity" may not be true considering the complexity of microbial communities and their complicate interactions with

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

confounding environmental factors or host properties. Thus it is worthy to engage nonlinear methods in reducing the dimension of metagenomic profiles.

In this paper, a nonlinear dimension reduction method– Isomap [21], [22] is proposed to analyze and visualize the large-scale protein family (Pfams) profiles from 45 ocean metagenomes in Global Ocean Sampling Expedition [23], [24]. To our knowledge, this is the first attempt to apply nonlinear methods in visualizing and analyzing metagenomic profiles. In this study, an interesting nonlinear relationships of Pfams are discovered by the method. Additionally, we provide biological interpretations of the nonlinear relationships by investigating the environmental variables of metagenome samples.

METHODS

Datasets

Pfam profile:: We used a same Pfam profile dataset to a previous work [19] which was derived from the metagenomes in Global Ocean Sampling expedition [23], [24]. We selected a subset of samples which had been processed in similar ways, in particular, we used only samples with filter size $0.1 - 0.8\mu m$, and excluded samples that appeared to represent completely distinct environmental conditions, such as those from freshwater environments, samples with very few reads and outliers in the a preliminary analysis. The final dataset is composed of 45 samples. Protein sequences from the unassembled reads for each sample were downloaded from CAMERA [25], and searched using HMMER 3.0 (http://hmmer.org) against protein families from the Pfam database version 24 [26] using Pfam's per-family gathering threshold cutoffs. We eliminated Pfams with fewer than 0.01 % of the total reads to avoid bias, leading to a read matrix with p =1584 rows and $s = 45$ samples, whose columns are normalized by dividing their sums respectively thus the column sums are equal to one.

Environmental factors and Geographic distance:: As we did in a previous work [19], we extracted six environmental factors. Salinity, sample depth, chlorophyll level, temperature and water depth are taken from the GOS metadata [23], [24], and total incident solar insolation at the surface was obtained from the NASA Surface meteorology and Solar Energy (SSE, http://eosweb.larc.nasa.gov/sse/) Release 6.0 Data Set. Missing environmental values are estimated as the average value for the respective variable. The square root of water depth was used in correlation analyses to eliminate over-weighting samples taken over the very deep ocean. Geographic distances were calculated as pairwise distances among sample locations using the great circle route as well as the latitude and longitude recorded in the GOS sample metadata. Geographic distances were log-transformed so as to not give undue weight to very large distances.

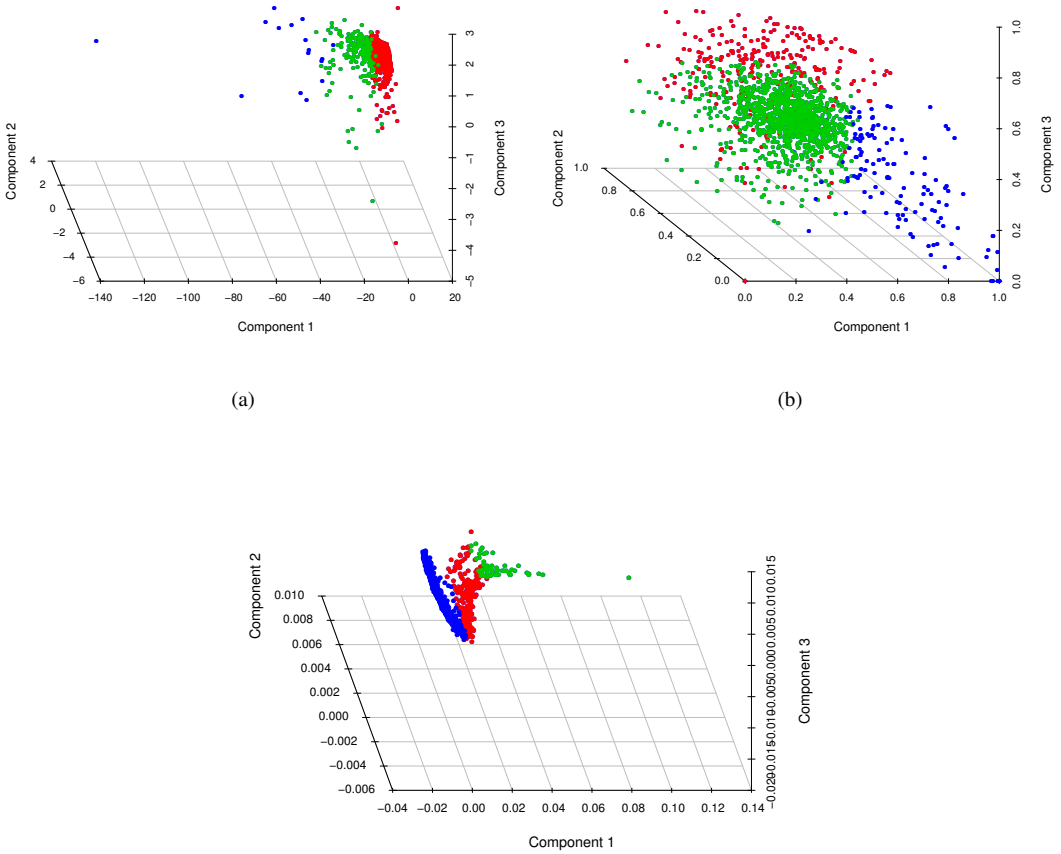
Isomap

Isomap is a global geometric framework for nonlinear dimensionality reduction [21] which is built on classical Multidimensional Scaling (MDS) [15]. Although MDS is developed to preserve the Eulidean distance in a low dimension space, Isomap seeks to preserve the intrinsic geometry of the data as captured in the geodesic manifold distances between all pairs of data points. If we have p Pfams and s samples, then the size of the profile matrix X is $p \times s$. Typically, Isomap has three steps [21]. In step 1, we firstly calculate the distance $d_X(i,j)$ between all pairs of points i, j in the input space X, where i, j could be the index of rows (Pfams) or columns (sites) depending on different study purposes. Then we determine which points are neighborhood on the manifold M depending on whether two points are close to each other within a distance threshold ϵ or one point is another's K nearest neighbors. In this study, we used the nearest neighbors method to determine the neighborhood of points on manifold ($K = 10$). A weighted graph with weight $d_x(x,j)$ over the data points is build to reconstruct the neighborhood relations in the original dataset. In step 2, the geodesic distances $d_M(i,j)$ are estimated by computing the shortest path distances $d_G(i,j)$ in the graph G. In step 3, the classical MDS [15] are applied on the matrix of graph distances $D_G = d_G i, j$. This step construct an embedding of the data in a d-dimensional Euclidean space Y that best preserves the manifold's estimated intrinsic geometry. The coordinate vectors y_i for points in Y are chosen to minimize the cost function $E =\| \tau(D_G) - \tau(D_Y) \|_{L^2}$, where D_Y denotes the matrix of Euclidean distances $d_Y(i,j) =\| y_i - y_j \|_{L^2}$ and $\| A \|_{L^2}$ denotes the L_2 norm of matrix A. The τ is a operator which converts distances to inner products which uniquely characterize the geometry of the data in a form that supports efficient optimization. The global minimum of the cost function is achieved by setting the coordinates y_i to the top d eigenvectors of the matrix τD_G. The true dimensionality of the data is estimated from the elbow plot which indicates the decreasing of the error as the dimensionality of Y is increased [21].

PCA

For PCA [12], we normalize the original Pfam profile by subtracting the means $\bar{X}_i$ from each x_i before constructing the covariance matrix so that each $\bar{X}_i$ has a mean of zero. This allows us to rewrite the covariance matrix as the following simple matrix multiplication: $\Sigma = \frac{1}{n} X X^T$. Then, by the spectral decomposition theory in linear algebra, we can decompose the matrix above into the product of three matrices: $\Sigma = U \Lambda U^T$, where $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ is the diagonal matrix of the eigenvalues of the covariance matrix ordered from highest to lowest.

Then, the principal components are the top d row vectors of U^T corresponding the d highest eigenvalues in Λ. We call $U^T$the projection weight matrix W and the transformed



(a) (b)

(c)

Figure 1. The organization structure of Pfams in a three dimensional space. (a) The top three principle components of PCA. (b) Three selected components of NMF. (c) The top three dimension after applying Isomap. The coloring is based on the k-means clustering on the top ten PCA and Isomap, and the three components of NMF.

data matrix S can be obtained from the original data matrix X by: $S = WX$. S is often called *scores* matrix, the rows of it are the low-dimension representatives of Pfams and the matrix W^T is called loadings matrix whose columns are of sample sites.

NMF

NMF decomposition finds matrices W and H, (with dimension $p \times k$ and $k \times s$, respectively, where k is the *rank* of our factorization) such that $WH \approx X$ [17]. We search for non-negative approximations that minimize the Kullback–Leibler (KL) divergence between X and WH [17]. The matrix W is $p \times k$ whereas the matrix H is $k \times s$. Hence, each column of W has one entry for each of the p Pfams; we can thus think of W as a collection of k "low-dimension samples", where k is the "degree" of the factorization. In this interpretation the s columns of H give each of the s environmental samples as linear combinations of these low-dimension samples. In the dual interpretation, the k rows of H are "low-dimension Pfams" and the p columns of W give the observed Pfams as linear combinations of them.We have introduced a method based on the H matrix for choosing an appropriate rank (k) for NMF analysis in the presence of overlap [18], [19]. We construct a symmetric similarity matrix $S = \hat{H}^T \hat{H}$, where $\hat{H}$ is column-normalized so that S has ones down the diagonal; thus each off-diagonal entry gives the similarity of two samples as seen by our NMF decomposition. We then define the "concordance index" $C = 1 - D$, where D is the mean squared difference between off-diagonal entries of S_j obtained from different realizations of the decomposition [18]. The concordance index C reflects the stability of this matrix across different

realizations of the factorization, and is used to select a good decomposition rank k.

Results and Discussion

Pattern inference

To understand the complex structure of microbial communities, we compared the results discovered by Isomap with those by PCA and NMF on a Pfam profile which characterize the distribution of 1584 protein families across 45 metagenomes of global ocean [23], [24]. PCA and NMF have been widely used in a broad field and have been recently adopted as methods to visualize and analyze metagenomic profiles. On this dataset, however, they all identified unclear clustering structures (Figure 1.a and b). In contrast with them, Isomap identified clear nonlinear structure embedded in the data (Figure 1.c). In Figure 1.a and b, Pfams are projected into the space spanned by the top three principle components of PCA and NMF respectively to visualize the Pfams. There are none good visualization of Pfams in the results of PCA and NMF. Figure 1.c indicate that Isomap can discover more interesting data structures than PCA and NMF. We colored the Pfams based on a k-means clustering on the $K = 10$ dimensions (we called components for convenience) of Isomap (see Figure 1.c). Interestingly, there is a Pfam group (blue in Figure 1.c) regardless of the the dimension one; however, the variation in another two groups (red and green group in Figure 1.c) are related to all the three dimensions. The biological meaning of these dimensions will be explored in the next section.

The driving environmental factor of dimensions

We investigate the following question: what are the major driving environmental factors of the top dimensions in Isomap? To understand the biological meaning of Isomap components in Figure 1.c, we compute the Pearson Correlation Coefficients (PCC) between the Pfam profiles of components and those of samples. A high PCC indicate a component has high similarity to a sample. Then we reorder the samples based on their correlation strengths to a component to see the trend of the associated environmental factors of a given component. We find that component 1 and component 2 are driven by distinctly different combination of environmental factors respectively. The first component has strong negative correlation with several environmental factors (salinity, temperature), and positive correlation with chlorophyll level (Table I). The second component has strong positive correlation with salinity, temperature and water depth, and negative correlation with chlorophyll level (see Table II). Analog to the correlations among NMF components and environmental variables [19], the results here indicate that the nonlinear components from Isomap have significant connection to environmental factors.

Table I
THE PEARSON CORRELATION COEFFICIENTS BETWEEN THE COMPONENT 1 OF ISOMAP AND ENVIRONMENTAL VARIABLES.

	PCC	p value
Salinity	-0.52	0.000
Sample_Depth	-0.23	0.133
Chlorophyll	0.46	0.002
Temperature	-0.43	0.003
Water_Depth	-0.08	0.587
Insolation	-0.36	0.016

Table II
THE PEARSON CORRELATION COEFFICIENTS BETWEEN THE COMPONENT 2 OF ISOMAP AND ENVIRONMENTAL VARIABLES.

	PCC	p value
Salinity	0.42	0.004
Sample_Depth	0.24	0.113
Chlorophyll	-0.58	0.000
Temperature	0.58	0.000
Water_Depth	0.48	0.001
Insolation	0.16	0.308

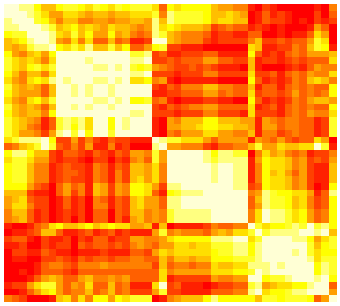


Figure 2. The similarity matrix of sites filtered by Isomap. The top three components are used to compute the similarity.

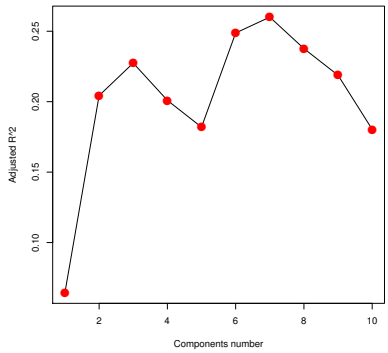


Figure 3. The correlations among Pfam profiles of top three components and samples.

Measuring functional distance in low dimensional space

Although it is direct to reduce the dimension of sample space, it is also useful to reduce that of Pfam space. The reduction of dimension of a Pfam space result in a low dimension Pfam profile $Y_{d\times s}$ where s denote the number of sites and d the number of dimension. We can compute the both similarity and distance matrix of s sites based on the d rows of Y. Based on the clear patterns in the site similarity matrix (Figure 2) of top 3 dimensions, we hypothesized that Isomap-filtered Pfam distance would be a useful metric for functional distance among sample sites. To test this idea, we compared how well different measures of functional distance were modeled by a combination of environmental distance and geographic distances in a naive regression model. Using adjusted R^2 as a measure of overall correlation, we see an interesting decrease of R^2 after 3 and 7 dimensions (Figure 3), indicating that adding some dimensions does not improves the model more than would be expected by chance. This may suggest the existence of noise in the Pfam profile and the dimension reduction can be viewed as a filter process of the noise. We also found that the correlation of Isomap-filtered Pfam distance with environmental and geographic distances (overall adjusted $R^2 = 0.26$) was comparable to that of PCA-filtered Pfam distance using the same number of components (adjusted $R^2 = 0.29$) and NMF-filtered distance (adjusted $R^2 = 0.32$), and higher than that of unfiltered Pfam distance (adjusted $R^2 = 0.22$). Note that adding a component does not mean the increasing of a adjusted R^2, the maximum of R^2 value is taken when top 1 to 10 components of PCA and Isomap are used to calculate R^2 respectively and that of rank 4 NMF (which is a local peak of concordance plot for model selection, data not shown). The result suggests that the Isomap filtering retains more information relevant to these correlations than direct distances but less information than NMF- and PCA-filtering. This is possibly because Isomap is a nonlinear method but we used a linear model to analyze their results. It may suggest that the employment of nonlinear statistical methods could be necessary in future investigation.

Conclusions

Interpretation of large-scale metagenomic datasets is technically challenging, but has the potential to provide important biological and biogeographic insights. Prior work has largely focused on assessing the distribution of taxonomic diversity and its relationship to environmental factors [27]–[29]. However, several recent studies have taken steps toward quantifying functional diversity of microbial communities and several authors have used GOS data to quantitatively investigate the relationship between microbial function and environment variables [13], [19], [20], [30].

The identification of complex structures and patterns of microbial communities is still at the essential part of studies in microbial ecology. Approaches based on linear assumption has been used extensively for discovering structures in metagenomic profiles and investigating the relationship between microbial function and metadata [13], [14], [19], [20], [30]. Unfortunately, metagenomic profile and their associated metadata are often complicated and they are nonlinearly interacted with each other. To overcome the challenge, we propose a analysis framework based on an nonlinear dimension reduction method–Isomap to explore the nonlinear structure in metagenomic profiles. Isomap is applied on the a protein family profile which are derived from the large-scale metagenomic dataset in GOS. The experimental results show that Isomap can discover novel and interesting data patterns which are not identified by other methods such as PCA and NMF.

The framework presented here is one of several approaches that can help shed light on the relationships among microbes [13], [14], [19], [20], [30]. The nonlinear method will be a necessary tool in metagenomics, not only because microbial function can be viewed at multi-scales: from individual genomes to communities to global cycles but also the complex interaction across scales. In future, it could be interesting to see the nonlinear correlations among microbial functions and environmental factors. Furthermore, it is also important to discover representative Pfam markers for a particular nonlinear dimension.

Acknowledgment

The authors would like to thank Jonathan Dushoff for helpful discussion, Morgan Langille for providing the Pfam profile and Russell Y. Neches for exploring the environmental variables and geographic distances. This work was supported in part by NSF CCF 0905291, NSF CCF 1049864, NSF IIP 1160960, NSFC 90920005, NSFC 61170189 and China National 12-5 plan 2012BAK24B01.

References

[1] J. Handelsman, "Metagenomics: application of genomics to uncultured microorganisms." *Microbiol Mol Biol Rev*, vol. 68, pp. 669–85, 2004 Dec.

[2] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes." *PLoS Biol*, vol. 5, p. e82, 2007 Mar.

[3] J. Raes and P. Bork, "Molecular eco-systems biology: towards an understanding of community function." *Nat Rev Microbiol*, vol. 6, pp. 693–9, 2008 Sep.

[4] P. J. Turnbaugh and J. I. Gordon, "An invitation to the marriage of metagenomics and metabolomics." *Cell*, vol. 134, pp. 708–13, 2008 Sep 5.

[5] E. A. Dinsdale *et al.*, "Functional metagenomic profiling of nine biomes." *Nature*, vol. 452, pp. 629–32, 2008 Apr 3.

[6] S. L. Strom, "Microbial ecology of ocean biogeochemistry: a community perspective." *Science*, vol. 320, pp. 1043–5, 2008 May 23.

[7] C. Quince *et al.*, "Accurate determination of microbial diversity from 454 pyrosequencing data." *Nat Methods*, vol. 6, pp. 639–41, 2009 Sep.

[8] J. Peterson *et al.*, "The nih human microbiome project." *Genome Res*, vol. 19, pp. 2317–23, 2009 Dec.

[9] J. Qin *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing." *Nature*, vol. 464, pp. 59–65, 2010 Mar 4.

[10] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 260–270, Mar. 2012. [Online]. Available: http://www.nature.com/nrg/journal/v13/n4/full/nrg3182.html

[11] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics." *PLoS Comput Biol*, vol. 6, p. e1000667, 2010 Feb.

[12] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics, Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, 2, no. 4, 4, pp. 433, 433–459, 459, Jul. 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/wics.101/abstract

[13] T. A. Gianoulis *et al.*, "Quantifying environmental adaptation of metabolic pathways in metagenomics." *Proc Natl Acad Sci U S A*, vol. 106, pp. 1374–9, 2009 Feb 3.

[14] M. Arumugam *et al.*, "Enterotypes of the human gut microbiome." *Nature*, vol. 473, pp. 174–80, 2011 May 12.

[15] S. S. Schiffman, *Introduction to multidimensional scaling : theory, methods, and applications /.* New York :: Academic Press,, 1981.

[16] M. Hamady, C. Lozupone, and R. Knight, "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data," *The ISME journal*, vol. 4, no. 1, pp. 17–27, Jan. 2010, PMID: 19710709 PMCID: PMC2797552.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–91, 1999 Oct 21.

[18] X. Jiang, J. Weitz, and J. Dushoff, "A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data," *Journal of Mathematical Biology*, vol. 64, no. 4, pp. 697–711, 2012. [Online]. Available: http://www.springerlink.com/content/t6t2536581304492/abstract/

[19] X. Jiang *et al.*, "Functional biogeography of ocean microbes revealed through non-negative matrix factorization," *PLoS One, Accepted*, 2012.

[20] J. Raes *et al.*, "Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data," *Molecular Systems Biology*, vol. 7, p. 473, Mar. 2011, PMID: 21407210 PMCID: PMC3094067.

[21] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000. [Online]. Available: http://www.sciencemag.org/content/290/5500/2319

[22] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, Jan. 2002. [Online]. Available: http://www.sciencemag.org/content/295/5552/7

[23] S. Yooseph *et al.*, "The sorcerer II global ocean sampling expedition: Expanding the universe of protein families," *PLoS Biology*, vol. 5, no. 3, p. e16, 2007. [Online]. Available: http://biology.plosjournals.org/perlserv/?request=get-document\&doi=10.1371%2Fjournal.pbio.0050016

[24] D. B. Rusch *et al.*, "The sorcerer II global ocean sampling expedition: Northwest atlantic through eastern tropical pacific," *PLoS Biology*, vol. 5, no. 3, p. e77, 2007.

[25] S. Sun *et al.*, "Community cyberinfrastructure for advanced microbial ecology research and analysis: the camera resource." 2011 Jan, vol. 39, pp. D546–51.

[26] R. D. Finn *et al.*, "The pfam protein families database." *Nucleic Acids Res*, vol. 38, pp. D211–22, 2010 Jan.

[27] J. L. Green, B. J. M. Bohannan, and R. J. Whitaker, "Microbial biogeography: From taxonomy to traits," *SCIENCE*, vol. 320, no. 5879, pp. 1039–1043, MAY 23 2008.

[28] S. L. Strom, "Microbial ecology of ocean biogeochemistry: A community perspective," *SCIENCE*, vol. 320, no. 5879, pp. 1043–1045, MAY 23 2008.

[29] J. A. Eisen, "Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes," *PLOS BIOLOGY*, vol. 5, no. 3, pp. 384–388, MAR 2007.

[30] P. V. Patel *et al.*, "Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families." *Genome Res*, vol. 20, pp. 960–71, 2010 Jul.