AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Biogeosciences

RESEARCH ARTICLE

10.1002/2014JG002676

Key Points:
- Method for estimating beta diversity with sequence data was developed
- Beta diversities between eight archaeal communities were estimated
- Model of archaeal beta diversity patterns was constructed

Correspondence to:
H. Koyano,
koyano@kuicr.kyoto-u.ac.jp

Citation:
Koyano, H., T. Tsubouchi, H. Kishino, and T. Akutsu (2014), Archaeal β diversity patterns under the seafloor along geochemical gradients, *J. Geophys. Res. Biogeosci.*, *119*, 1770–1788, doi:10.1002/2014JG002676.

Received 24 MAR 2014
Accepted 30 JUL 2014
Accepted article online 12 AUG 2014
Published online 2 SEP 2014

Archaeal β diversity patterns under the seafloor along geochemical gradients

Hitoshi Koyano[1], Taishi Tsubouchi[2], Hirohisa Kishino[3], and Tatsuya Akutsu[1]

[1]Institute for Chemical Research, Kyoto University, Uji, Japan, [2]Japan Agency for Marine–Earth Science and Technology, Yokosuka, Japan, [3]Graduate School of Agricultural and Life Sciences, University of Tokyo, Bunkyo, Japan

Abstract Recently, deep drilling into the seafloor has revealed that there are vast sedimentary ecosystems of diverse microorganisms, particularly archaea, in subsurface areas. We investigated the β diversity patterns of archaeal communities in sediment layers under the seafloor and their determinants. This study was accomplished by analyzing large environmental samples of 16S ribosomal RNA gene sequences and various geochemical data collected from a sediment core of 365.3 m, obtained by drilling into the seafloor off the east coast of the Shimokita Peninsula. To extract the maximum amount of information from these environmental samples, we first developed a method for measuring β diversity using sequence data by applying probability theory on a set of strings developed by two of the authors in a previous publication. We introduced an index of β diversity between sequence populations from which the sequence data were sampled. We then constructed an estimator of the β diversity index based on the sequence data and demonstrated that it converges to the β diversity index between sequence populations with probability of 1 as the number of sampled sequences increases. Next, we applied this new method to quantify β diversities between archaeal sequence populations under the seafloor and constructed a quantitative model of the estimated β diversity patterns. Nearly 90% of the variation in the archaeal β diversity was explained by a model that included as variables the differences in the abundances of chlorine, iodine, and carbon between the sediment layers.

1. Introduction

In recent years, advancements in geobiology have revealed that diverse microorganisms, particularly archaea, live in subsurface areas [*Parkes et al.*, 1994; *Whitman et al.*, 1998; *D'Hondt et al.*, 2004; *Lipp et al.*, 2008]. In this study, we investigated β diversity patterns and their determinants in archaeal communities in layers below the seafloor.

The extent to which two biological communities differ is called β diversity [*Whittaker*, 1960, 1972]. In ecology, many studies have been conducted on methods for measuring β diversity, which has traditionally been measured using presence-absence data of species. Numerous methods for measuring β diversity with presence-absence data of species have been developed to date [*Jaccard*, 1900; *Dice*, 1945; *Sørensen*, 1948; *Ochiai*, 1957; *Whittaker*, 1960; *Anderberg*, 1973; *Cody*, 1975; *Routledge*, 1977; *Wilson and Shmida*, 1984; *Harrison et al.*, 1992; *Cody*, 1993; *Colwell and Coddington*, 1994; *Weiher and Boylen*, 1994; *Lande*, 1996; *Williams*, 1996; *Mourelle and Ezcurra*, 1997; *Harte and Kinzig*, 1997; *Ruggiero et al.*, 1998; *Williams et al.*, 1999; *Gaston et al.*, 2001; *Lennon et al.*, 2001]. See, for example, *Southwood and Henderson* [2000], *Koleff et al.* [2003], and *Magurran* [2004] for a review. Typically, samples do not represent complete lists of the species in the two biological communities (i.e., the populations from which the samples were collected). Therefore, the measurement of β diversity is actually an estimation of the β diversity between populations based on the samples collected. However, the methods listed above for measuring β diversity do not distinguish between a population and a sample. They only introduce an index of β diversity between two samples without defining a β diversity index between two populations. Consequently, these methods do not evaluate the accuracy with which the β diversity index calculated based on samples estimates the β diversity between populations that is the quantity of interest, where the term accuracy was used in the statistical sense to represent the extent to which an estimator based on random samples from populations could accurately estimate the β diversity between these populations. Studies on methods that make a distinction between a population and a sample for measuring β diversity using presence-absence data of species include *Smith et al.* [1996], *Plotkin and Muller-Landau* [2002], and *Chao et al.* [2005]. Among these studies, only *Plotkin and*

Muller-Landau [2002] theoretically examined the accuracy of their estimator. These authors showed that their estimator was unbiased in a parametric framework.

Methods for measuring β diversity have been developed that account for species frequencies as well as the presence or absence of individual species. This approach avoids considering two communities as identical when their species compositions are the same but the frequencies of each species differ between them. Indices of β diversity that consider species frequencies include the Bray-Curtis index [*Czekanowski*, 1909; *Renkonen*, 1938; *Motyka*, 1947; *Odum*, 1950; *Bray and Curtis*, 1957], the Morishita-Horn index [*Morisita*, 1959; *Horn*, 1966], and the Gower index [*Gower*, 1971; *Anderson et al.*, 2006] (these indices were constructed for application to species data, but they can also be applied to sequence data). These indices do not distinguish between a population and a sample. In contrast, the Yue-Clayton index [*Yue and Clayton*, 2005] is a measure of β diversity that accounts for the frequencies of species and distinguishes between a population and a sample, although it has not been theoretically examined how accurately this index calculated based on samples estimates the β diversity index between populations.

None of the indices listed above consider the divergence between species. As described in *Lozupone and Knight* [2008], methods for measuring β diversity have progressed from those that do not consider the divergence between species or sequences to those that do. These methods have been developed because when a species is present in community A but not community B, it is relevant to consider whether all species in community B are distantly related to the focal species in community A or whether a closely related species is present (see *Lozupone and Knight* [2008] for the advantages of considering divergence). The taxonomic (dis)similarity proposed by *Izsak and Price* [2001] is the first measure of β diversity that accounts for the divergences between species. Species are not clearly defined for microorganisms, unlike for animals and plants; therefore, sequence data (for example, environmental samples of 16S ribosomal RNA gene sequences) are used to measure microbial diversity. It is more important to distinguish between the above two cases in measuring the β diversity with sequence data than with species data because the exact same sequences are rarely collected in two different environments. Indices of β diversity that use sequence data and consider the divergence between sequences include the unique fraction metric (UniFrac) [*Lozupone and Knight*, 2005] (see also *Lozupone and Knight* [2008]). The Izsak and Price (dis)similarity and UniFrac are qualitative measures of β diversity that do not account for the frequencies of species or sequences. Furthermore, neither of these measures distinguishes between a population and a sample.

Weighted UniFrac [*Lozupone et al.*, 2007], which is an extension of UniFrac, and double principal coordinates analysis (DPCoA) [*Pavoine et al.*, 2004], which employs Rao's dissimilarity [*Rao*, 1982], are β diversity measures that use sequence data and account for both the frequency of each sequence and the divergence between the sequences. However, these methods were not designed in the framework of estimating β diversity between two populations based on the samples drawn from the populations. Recently, a large amount of sequence data have become available, but an environmental sample of biological sequences is a small part of the population of all of the sequences in one environment, especially for microbial communities. Thus, desirable methods are those designed in the statistical framework in which a distinction between a population and a sample is made, an index of the β diversity between populations is first defined, and an estimator based on samples for the index is subsequently constructed. Furthermore, the estimator must be demonstrated to accurately estimate the β diversity index between populations.

In conclusion, for the analysis of β diversity between microbial communities, the following criteria for a method for measuring β diversity are required: (i) It uses sequence data, not species data. (ii) It considers both the frequency of each sequence and the divergence between the sequences. (iii) It is constructed in the statistical framework of estimating the β diversity between populations based on samples. And (iv) it is theoretically demonstrated that the estimator based on samples accurately estimates the β diversity index between populations. Therefore, in this study, we systematically addressed the problem of estimating β diversity with sequence data by applying probability theory on a set of strings that two of the authors developed in a previous publication [*Koyano and Kishino*, 2010]. We first defined an index of β diversity between populations of sequences as a distance that reflects the frequencies of sequences and the divergences between sequences. Then, we constructed an estimator of this β diversity index and demonstrated that the estimator has the property of strong consistency. A detailed description of our method is provided in section 3.

Many analyses of the β diversity patterns of various biological communities with respect to different environmental variables and gradients have been conducted to date. Environmental variables and gradients include (i) latitude [*Rodriguez and Arita*, 2004; *Qian and Ricklefs*, 2007], (ii) altitude [*Brehm et al.*, 2003], (iii) sea depth [*Izsak and Price*, 2001], (iv) temperature [*Miller et al.*, 2009], (v) salinity [*Walsh et al.*, 2005; *Santoro et al.*, 2006], (vi) areas [*Harrison et al.*, 1992; *Condit et al.*, 2002], and (vii) internal organs and skin of subjects [*Eckburg et al.*, 2005; *Gao et al.*, 2007]. In this study, we sought environmental variables that are systematic factors for variation in the β diversity of archaeal communities in layers below the seafloor. We first investigated differences between layers in the abundances of different elements and compounds as candidates for such environmental variables, evaluating whether there were quantitative relationships involving the β diversities between archaeal communities and the differences between their environments. As explained in the fourth paragraph of section 5, we did not use composite variables, such as (i), (ii), (iii), (vi), and (vii) among the seven listed above, as environmental variables to model archaeal β diversity patterns. We next modeled archaeal β diversity patterns by developing a numerical equation including these environmental variables.

We assume that $\nu \geq 2$. C_i represents a set of all biological sequences (for example, 16S ribosomal RNA gene sequences) that corresponds to a biological community that was set as the object of analysis in the ith environment for each $i = 1, \cdots, \nu$. Let $d_\beta\left(C_i, C_{i'}\right)$ denote a β diversity index between C_i and $C_{i'}$, which is precisely defined in section 3. Let $v_1, \cdots, v_\kappa$ be real-valued environmental variables, which are systematic factors for the variation in $d_\beta\left(C_i, C_{i'}\right)$. We denote a measurement of v_j in the ith environment by v_{ji} for each $j = 1, \cdots, \kappa$ and $i = 1, \cdots, \nu$. Examples of v_j include temperature, hydrogen-ion concentration, and other factors. β diversity is a type of distance between two communities, and the distance between two real numbers is the absolute value of the difference between them. Therefore, in this setting, the problem described above is to make a list of environmental variables and to find a function $\varphi : [0, \infty)^\kappa \to [0, \infty)$ such that

$$d_\beta\left(C_i, C_{i'}\right) = \varphi\left(|v_{1i} - v_{1i'}|, \cdots, |v_{\kappa i} - v_{\kappa i'}|\right) + \zeta_{ii'}, i = 1, \cdots, \nu - 1, i' = i + 1, \cdots, \nu \quad (1)$$

for an error term $\zeta_{ii'}$. However, $C_1, \cdots, C_\nu$ are generally not available, as described above. Let S_i be a sample of sequences collected from C_i. We denote an estimator of $d_\beta\left(C_i, C_{i'}\right)$ based on S_i and $S_{i'}$ by $\widehat{d}_\beta\left(S_i, S_{i'}\right)$. Then, equation (1) is modified to

$$\widehat{d}_\beta\left(S_i, S_{i'}\right) = \varphi\left(|v_{1i} - v_{1i'}|, \cdots, |v_{\kappa i} - v_{\kappa i'}|\right) + \varepsilon_{ii'}, i = 1, \cdots, \nu - 1, i' = i + 1, \cdots, \nu, \quad (2)$$

where $\varepsilon_{ii'}$ is the sum of the error term $\zeta_{ii'}$ in equation (1) and the estimation error of $d_\beta\left(C_i, C_{i'}\right)$ by $\widehat{d}_\beta\left(S_i, S_{i'}\right)$. It would be a reasonable approach to seek an appropriate function φ in the set of linear functions first. In this case, our problem is to make a list of environmental variables $v_1, \cdots, v_\kappa$ and to determine constants $b_0, \cdots, b_\kappa > 0$ such that

$$\widehat{d}_\beta\left(S_i, S_{i'}\right) = b_0 + b_1|v_{1i} - v_{1i'}| + \cdots + b_\kappa|v_{\kappa i} - v_{\kappa i'}| + \varepsilon_{ii'}, i = 1, \cdots, \nu - 1, i' = i + 1, \cdots, \nu$$

(note the sign condition for $b_0, \cdots, b_\kappa$). We address this problem in section 4. To the authors' knowledge, no quantitative model of β diversity patterns has been reported that is expressed as a numerical equation with respect to the environmental variables.

Finally, we describe future challenges in modeling the β diversity patterns along geochemical gradients using the quantitative approach taken in this study and a new problem raised by the results of this study in section 5.

2. Materials

In this study, we analyzed data that were collected from a sediment core of 365.3 m that was obtained by drilling into the seafloor off the east coast of the Shimokita Peninsula during the second shakedown cruise of D/V *Chikyu* (August to October 2006) in the Integrated Ocean Drilling Program that started in October 2003. In this section, we briefly describe the materials analyzed in the following sections.

2.1. Study Site and a Core Sample

The drilling site (41°10'38.28"N–142°12'04.89"E) was located in the Sanriku-oki sedimentary basin, a northeastern forearc basin of the Japanese main island [see *Tomaru et al.*, 2009, Figures 1A and 1B]. This site was 1180 m in water depth and was drilled to a depth of 365 m below the seafloor (mbsf). There were gas

hydrates in the area around the study site. An abundance of natural gas was found in the Paleocene to Eocene sequences in the Ministry of International Trade and Industry (MITI) Sanriku-oki well near the site. In the Sanriku-oki, the Pacific Plate is subducting beneath the North American Plate along the Japan Trench. The rate of subduction was estimated at up to 9 cm/yr [*von Huene and Culotta*, 1989]. The Pacific Plate in this region is the oldest of currently subducting plates, with an age of 140 Ma [*Nakanishi et al.*, 1989, 1992]. The Shimokita area is of great geological interest. See *Taira et al.* [2005] and *Tomaru et al.* [2009] for a survey of this area.

A core from the drilled hole had a highly continuous sequence. Sediments collected from the core were mainly composed of silty clay, and ash layers were frequently intercalated between silt and sand layers. Between 30 and 40% of the silty clay elements comprised biogenic particles, and large amounts of quartz and clay minerals were found. Developed lamination was rare, and many microfossils were observed. These fossils were dominated by diatoms and included siliceous sponge spicules, planktonic foraminiferans, benthic foraminiferans, dinoflagellates, and radiolarians. From the integrated analysis based on micropaleontology, tephrochronology, and magnetostratigraphy, the bottom age of the hole at 365 mbsf was estimated at 780 ka. The rate of sedimentation at the hole was estimated as nearly constant at approximately 62 cm/kyr.

2.2. Geochemical and Archaeal Sequence Data

After recovery of the core, 10–15 cm sections of sediment were sampled from core sections and immediately skinned to avoid a potential source of contamination. Physical and chemical data were then measured onboard. For example, pore water samples were collected from the sections with a Manheim-type squeezing system [*Manheim et al.*, 1994], and the dissolved SO_4 and Cl concentrations were measured from aliquots of the water samples using ion chromatography. The measurement of the total dissolved Br and I concentrations was conducted using inductively coupled plasma mass spectrometry. The processes of recovering cores and measuring various data were described previously in detail [*Aoike*, 2007]. See *Tomaru et al.* [2009] for the geochemical data used in this study.

Furthermore, microbial DNA was extracted from the core from the eight layers at depths of 0.7, 4.9, 11.0, 18.5, 48.0, 107.0, 217.0, and 348.5 mbsf. A detailed description was given in *Nunoura et al.* [2005] on sampling, DNA isolation, and fosmid library construction. The procedures for screening for microbial genome fragments encoding small subunit ribosomal RNA genes, sequencing and enrichment of the fragments, and annotation have been described previously [*Nunoura et al.*, 2011]. Archaeal sequences in the constructed library of bacterial and archaeal 16S ribosomal RNA gene sequences were analyzed in the present study. The numbers of gene sequences analyzed from the first to eighth layers were 1411, 1075, 1121, 1360, 1291, 1062, 1178, and 1267.

3. Method for Estimating β Diversity

In this section, we treat the problem of measuring β diversity with sequence data based on probability theory on a set of strings in a rigorous manner. As described in section 1, the measurement of β diversity is actually an estimation of β diversity between populations based on samples. Sequence populations that can be addressed using the framework described below include the sets of all 16S ribosomal RNA gene sequences in two environments because the Levenshtein distance [*Levenshtein*, 1966] was used as the distance between two sequences (the Levenshtein distance between two sequences s and t is the minimal number of deletions, insertions, or substitutions required to transform s into t). We first define a β diversity index between sequence populations in two environments in section 3.1. We must estimate the introduced β diversity index between sequence populations based on samples because it is physically impossible to collect all sequences (for example, all microbial 16S ribosomal RNA gene sequences) in two environments. Therefore, we construct an estimator of the β diversity index based on sequence data and demonstrate that the estimator converges to the index between sequence populations with probability of 1 as the number of collected sequences increases in section 3.2.

3.1. Formulation of the β Diversity Index Between Sequence Populations

In this subsection, we consider the problem of defining the β diversity index between populations of sequences. As described in section 1, in this study, we define the β diversity index as a distance that reflects the frequencies of sequences and the divergences between sequences. In the following paragraph, we introduce the β diversity index between sequence populations as a distance between population distributions of

sequences in two steps by extending the distance between probability mass or density functions, which has been used in the mathematical fields of functional analysis and probability theory. $\mathbb{Z}^+$, $\mathbb{N}$, and $\mathbb{R}$ represent the sets of positive integers, natural numbers (including 0), and real numbers, respectively.

Let $\mathcal{P}$ be the set of probability density functions defined on $\mathbb{R}$, that is,

$$\mathcal{P} = \left\{ f : \mathbb{R} \to [0, \infty) : \int_{\mathbb{R}} f(x)\mathrm{d}x = 1 \right\}.$$

Generally, the distance space $(\mathcal{P}, d_p)$ is constructed with the distance d_p defined by

$$d_p(f, g) = \left\{ \int_{\mathbb{R}} |f(x) - g(x)|^p \mathrm{d}x \right\}^{1/p} \tag{3}$$

for $p \in \mathbb{Z}^+$ (in many cases, $p = 2$) [see, for example, *Shiryaev*, 1996]. We first introduce a distance between sequence populations that reflects the frequencies of sequences by constructing an analogy to the distance (3) on a set of distributions of random strings. We use the framework of probability theory on a set of strings that was proposed in *Koyano and Kishino* [2010]. Several definitions necessary in the following text are cited in section A1. A random string σ is introduced as a type of a discrete stochastic process that takes values in a set A^* of sequences of letters in an alphabet (that is, a finite set of letters) A. We denote a set of random strings by $\mathcal{M}(\Omega, A^*)$. The length of σ, denoted by $|\sigma|$, and the probability function of the finite-dimensional distribution of σ at sites $i_1, \cdots, i_k \in \mathbb{Z}^+$, represented by $\boldsymbol{q}_{\sigma;i_1,\cdots,i_k}$, are defined, and the independence of a sequence of random strings $\{\sigma_n : n \in \mathbb{Z}^+\}$ is formulated.

For the probability function $\boldsymbol{q}_{\sigma;1,\cdots,|\sigma|}$ of the finite-dimensional distribution at sites $1, \cdots, |\sigma|$ of $\sigma \in \mathcal{M}(\Omega, A^*)$, we define the function $\boldsymbol{q}_\sigma : A^* \to [0, 1]$ as

$$\boldsymbol{q}_\sigma(s) = \begin{cases} \boldsymbol{q}_{\sigma;1,\cdots,|\sigma|}(x_1, \cdots, x_{|\sigma|}) & (\text{for } x_1, \cdots, x_{|\sigma|} \in \bar{A} \text{ such that} \\ & \quad s = (x_1, \cdots, x_{|\sigma|}, e, \cdots) \text{ if } |\sigma| \geq |s|) \\ 0 & (\text{if } |\sigma| < |s|). \end{cases}$$

$\boldsymbol{q}_\sigma$ is a probability function on A^*. We set

$$\mathcal{Q} = \{\boldsymbol{q}_\sigma : \sigma \in \mathcal{M}(\Omega, A^*)\}.$$

We introduce an analogy of the distance (3) on $\mathcal{Q}$.

Definition 1: We define the function $d_\beta : \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ as

$$d_\beta(\boldsymbol{p}, \boldsymbol{q}) = \sum_{s \in A^*} |\boldsymbol{p}(s) - \boldsymbol{q}(s)|. \tag{4}$$

It is easily verified that d_β is a distance on $\mathcal{Q}$.

We can add the operations of the pth power and pth root to the right-hand side of equation (4) of d_β. However, d_β defined in this manner does not have especially good properties when $p = 2$, unlike in the space of integrable functions. Therefore, we defined d_β as an analogy of the distance (3) with $p = 1$.

We have introduced the distance between sequence populations that reflects the frequencies of sequences in Definition 1. We next extend distance (4) such that it reflects the divergences between the sequences. We set

$$D_{\boldsymbol{q}_\sigma} = \{s \in A^* : \boldsymbol{q}_\sigma(s) > 0\}$$

for $\boldsymbol{q}_\sigma \in \mathcal{Q}$. $D_{\boldsymbol{q}_\sigma}$ is the support of $\boldsymbol{q}_\sigma$. We extend distance (4) as follows:

Definition 2: We set the function $\delta_{\boldsymbol{p},\boldsymbol{q}} : A^* \to [1, \infty)$ to be

$$\delta_{\boldsymbol{p},\boldsymbol{q}}(s) = \begin{cases} 1 & (\text{if } s \in D_{\boldsymbol{p}} \cap D_{\boldsymbol{q}}) \\ \min\{d_L(s, t) : t \in D_{\boldsymbol{q}}\} + 1 & (\text{if } s \in D_{\boldsymbol{p}} \text{ and } s \notin D_{\boldsymbol{q}}) \\ \min\{d_L(s, t) : t \in D_{\boldsymbol{p}}\} + 1 & (\text{if } s \notin D_{\boldsymbol{p}} \text{ and } s \in D_{\boldsymbol{q}}) \\ \text{an arbitrary real number} \geq 1 & (\text{if } s \notin D_{\boldsymbol{p}} \cup D_{\boldsymbol{q}}) \end{cases} \tag{5}$$

for $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{Q}$, and then, we redefine the function $d_\beta : \mathcal{Q} \times \mathcal{Q} \to [0, \infty)$ as

$$d_\beta(\boldsymbol{p}, \boldsymbol{q}) = \sum_{s \in A^*} \delta_{\boldsymbol{p},\boldsymbol{q}}(s) |\boldsymbol{p}(s) - \boldsymbol{q}(s)|. \tag{6}$$

It is easily verified that d_β is a distance on $\mathcal{Q}$. We define the β diversity between populations of sequences as this distance d_β.

If an index of β diversity between sequence populations that is normalized on the interval $[0, 1]$ is needed, then

$$d'_\beta(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{4} \sum_{s \in A^*} \delta'_{\boldsymbol{p},\boldsymbol{q}}(s) |\boldsymbol{p}(s) - \boldsymbol{q}(s)|$$

is useful, where $\delta'_{\boldsymbol{p},\boldsymbol{q}}(s)$ is a weight obtained by replacing the Levenshtein distance $d_L(s, t)$ in equation (5) with the Levenshtein distance per site $d'_L(s, t) = d_L(s, t)/\max\{|s|, |t|\}$.

3.2. Proposed Estimator of the β Diversity Index and Its Strong Consistency

In this subsection, we construct an estimator of the β diversity index (6) between populations of sequences introduced in the previous subsection, and we demonstrate that this estimator (described in equation (7) below) has the property of strong consistency (i.e., the estimator (7) based on samples converges to the β diversity index (6) between populations with probability of 1 as the sample sizes increase). Let $\{\sigma_1, \cdots, \sigma_m\}$ and $\{\tau_1, \cdots, \tau_n\}$ be sequences of independent random strings that have the identical distributions $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{Q}$, respectively. We denote realizations of σ_i for each $i = 1, \cdots, m$ and τ_j for each $j = 1, \cdots, n$ by s_i and t_j, respectively, and set $S^{(m)} = \{s_1, \cdots, s_m\}$ and $T^{(n)} = \{t_1, \cdots, t_n\}$. We formulate the problem of measuring the β diversity between sequence populations using sequence data as the problem of estimating $d_\beta(\boldsymbol{p}, \boldsymbol{q})$ based on $S^{(m)}$ and $T^{(n)}$. We estimate $\delta_{\boldsymbol{p},\boldsymbol{q}}(s)$ by

$$\hat{\delta}_{S^{(m)},T^{(n)}}(s) = \begin{cases} 1 & (\text{if } s \in S^{(m)} \cap T^{(n)}) \\ \min\{d_L(s, t) : t \in T^{(n)}\} + 1 & (\text{if } s \in S^{(m)} \text{ and } s \notin T^{(n)}) \\ \min\{d_L(s, t) : t \in S^{(m)}\} + 1 & (\text{if } s \notin S^{(m)} \text{ and } s \in T^{(n)}) \\ \text{an arbitrary real number} \geq 1 & (\text{if } s \notin S^{(m)} \cup T^{(n)}) \end{cases}$$

and then $d_\beta(\boldsymbol{p}, \boldsymbol{q})$ by

$$\hat{d}_\beta(\boldsymbol{p}, \boldsymbol{q}) = \sum_{s \in S^{(m)} \cup T^{(n)}} \hat{\delta}_{S^{(m)},T^{(n)}}(s) |\hat{\boldsymbol{p}}_{S^{(m)}}(s) - \hat{\boldsymbol{q}}_{T^{(n)}}(s)|, \tag{7}$$

where $\hat{\boldsymbol{p}}_{S^{(m)}}(s)$ and $\hat{\boldsymbol{q}}_{T^{(n)}}(s)$ represent the relative frequencies of $s \in A^*$ in $S^{(m)}$ and $T^{(n)}$, respectively. The following result can be obtained regarding the accuracy of estimator (7):

Proposition 1: In the above setting, $\hat{d}_\beta(\boldsymbol{p}, \boldsymbol{q})$ is a strongly consistent estimator of $d_\beta(\boldsymbol{p}, \boldsymbol{q})$ for any $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{Q}$.

The proof of this proposition is provided in section A2.

4. Results

In this section, we first examine the usefulness of our proposed method in practical data analysis by applying it to estimate the β diversities between the archaeal communities in the eight layers below the seafloor off the east coast of the Shimokita Peninsula. We subsequently construct a quantitative model of the archaeal β diversity patterns, in which several geochemical variables, such as the abundances of carbon and chlorine in the layers, are used as the environmental variables.

4.1. Estimation of the Archaeal β Diversities

In this subsection, we estimate the β diversities between the eight archaeal communities by applying the method proposed in the previous section and existing methods to the environmental samples of 16S ribosomal RNA gene sequences collected from these communities and then compare the results. Material flux is expected to exist in the sedimentary layers where the samples were collected [see, for example, *Elderfield et al.*, 1999; *D'Hondt et al.*, 2004]. Therefore, the archaeal communities in the close layers are inferred to be similar, and the β diversity between these communities would be small. (i) Therefore, if an estimating method produces estimates that are remarkably inconsistent with the order of layers of β diversities between the

communities, the method likely has some problems. The eighth layer is considerably deeper than the other seven, as described in section 2.2. (ii) Thus, for the same reason, a method that estimates that the β diversities between the communities in the eighth and other seven layers are less than those in the first to seventh layers would be considered inaccurate. Furthermore, the first layer is a subsurface layer, and consequently, a biological disturbance would more frequently occur by currents and other factors in the first layer than in the other seven. (iii) Therefore, a method that estimates that the β diversities between the communities in the first and other seven layers are less than those in the second to eighth layers is thought to be problematic. In the following paragraph, we examine estimation results using these three criteria.

We estimated the β diversities between the archaeal communities in the eight layers using the Sørensen index [*Dice*, 1945; *Sørensen*, 1948], the Bray-Curtis index [*Bray and Curtis*, 1957], and the Morishita-Horn index [*Morisita*, 1959; *Horn*, 1966], along with our proposed estimator, and then, we applied multidimensional scaling (MDS) [*Torgerson*, 1952] to the estimated β diversities to position the eight communities on the plane (Figures 1a–1d). Translation and scale transformation were applied to the two-dimensional coordinates that were obtained by MDS to facilitate the comparisons. The Sørensen index does not account for both the frequencies of the sequences and the divergences between sequences, whereas the Bray-Curtis and Morishita-Horn indices consider the frequencies of the sequences but do not consider the divergences between sequences. First, Figure 1a presents the results from the Sørensen index. In this figure, the communities in the first and eighth layers were located near to those in the second and sixth layers, respectively, and the communities in the first and eighth layers were not separated from those in the other six layers. In addition, the distances from the community in the eighth layer to those in the first, second, and fourth layers were almost equal to the distances from the communities in the fifth and sixth layers to those in the first, second, and fourth layers, respectively. Next, the results from the Bray-Curtis index, which are shown in Figure 1b, were similar to those obtained from the Sørensen index. The communities in the first and eighth layers were not separated from those in the other layers, and the locations of the communities were inconsistent with the order of layers. In Figure 1c, which presents the results from the Morishita-Horn index, the community in the first layer was not separated from those in the other seven layers, and the inconsistency between the locations of the communities and the order of layers was remarkable. Moreover, the distances from the community in the eighth layer to those in the first, second, and fourth layers were less than the distances from the communities in the third and sixth layers to those in the first, second, and fourth layers, respectively. Finally, the results from our method are provided in Figure 1d. In this figure, the communities in the first and eighth layers were located far from those in the other layers, and the locations of the eight communities were more consistent with the order of layers than in the results from the other methods. The communities in the second, third, and fourth layers were located in the same neighborhood in the result from our method, unlike from the other methods. These three layers can be interpreted as the limit of the movement of microorganisms by flux.

As described in section 1, not considering the frequencies of the sequences implies the identification of two communities as long as a list of the types of sequences in one community is equal to that of the other, even if the frequency distributions of sequences are different between the two communities. Not considering the divergences between sequences implies that when sequence S belongs to community A but not to community B, we identify the case in which community B includes sequences that are very similar to sequence S and the case in which any sequence in community B is different from sequence S. In the previous section, we introduced the new index of β diversity between sequence populations to handle these problems, and we theoretically demonstrated that our proposed estimator based on sequence data could accurately estimate the β diversity index. From the above results, our method appeared to be useful in the estimation of the β diversity between microbial communities with sequence data.

4.2. Modeling the Archaeal β Diversity Patterns Along Geochemical Gradients

In this subsection, we construct a quantitative model of the archaeal β diversity patterns under the seafloor along geochemical gradients by using the archaeal β diversities estimated in the previous subsection and data on the abundances of several elements and compounds in the eight layers below the seafloor. We first listed candidates for independent variables of the model on the basis of Pearson's correlation coefficients. The β diversity between two communities is a type of distance between them. Therefore, correlation coefficients must be calculated between the β diversities and the absolute values of the differences between the layers in the abundances of elements or compounds because the abundance of an element or compound is expressed as a real number, and the distance between two real numbers is the absolute value of the dif-

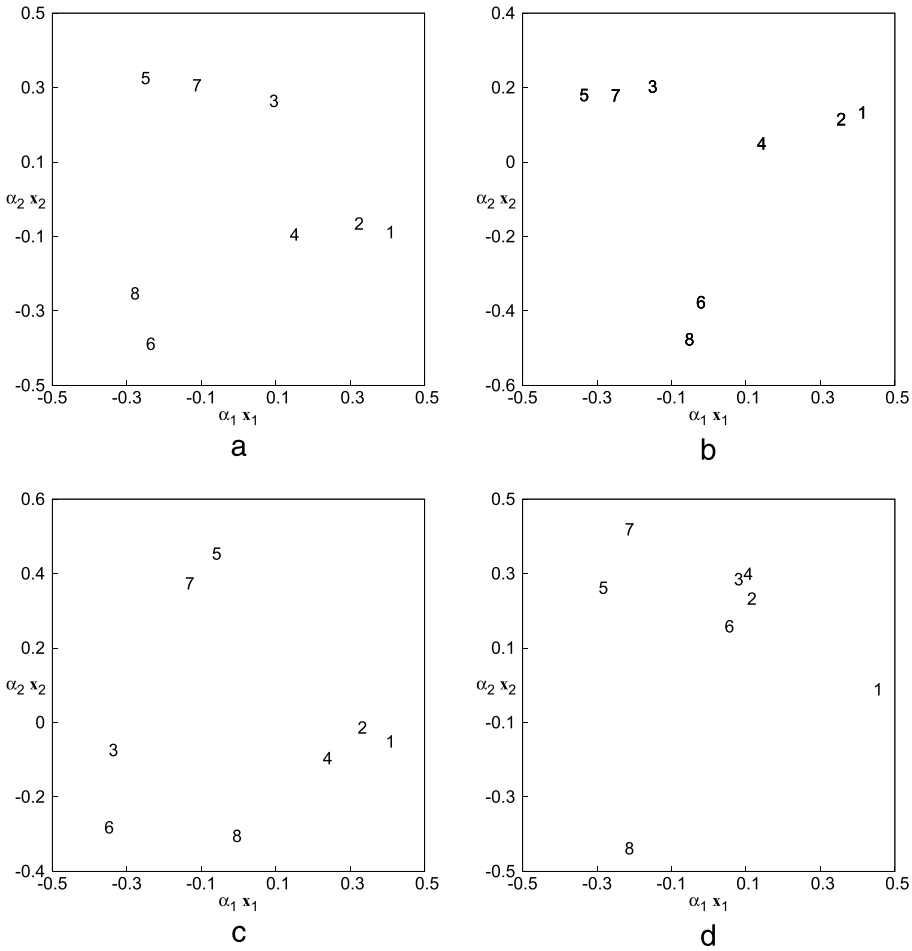

Figure 1. Results of applying MDS to the β diversities between the archaeal communities in the eight layers below the seafloor. Each number in the figure represents the layer number (see section 2). (a) The Sørensen index, (b) the Bray-Curtis index, (c) the Morishita-Horn index, and (d) the β diversity index proposed in this study. Here α_1 and α_2 are the greatest and second-greatest eigenvalues of the matrix $Y = (y_{ij})$, respectively, and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the eigenvectors for α_1 and α_2, respectively, where $y_{ij} = -\left(\beta_{ij}^2 - \sum_{j=1}^{8}\beta_{ij}^2/8 - \sum_{i=1}^{8}\beta_{ij}^2/8 + \sum_{i=1}^{8}\sum_{j=1}^{8}\beta_{ij}^2/8^2\right)/2$, and β_{ij} represents the β diversity between archaeal communities in the ith and jth layers for each $i, j = 1, \cdots, 8$.

ference between them. Table 1 presents the Pearson's correlation coefficients between the estimates of β diversity from our proposed method between the archaeal communities and the absolute values of the differences in the concentrations of SO_4^{2-} (mM), Cl^- (mM), Br^+ (μm), I (μM), methane (μM), bulk-CN sulfur (wt %), bulk-CN carbon (wt %), and bulk-CN nitrogen (wt %) between the eight layers. Henceforth, bulk-CN sulfur, bulk-CN carbon, and bulk-CN nitrogen are abbreviated as S, C, and N, respectively. From this table, we can see that the β diversity pattern of the archaeal communities has intermediate correlations with I, C, and N and a strong correlation with Cl^-.

For comparison, we calculated Pearson's correlation coefficients between each of the other β diversity indices (the *Anderberg* [1973], Bray-Curtis [*Czekanowski*, 1909; *Bray and Curtis*, 1957], *Jaccard* [1900],

Table 1. Pearson's Correlation Coefficients of the Estimated β Diversities Between the Archaeal Communities in the Eight Layers and the Absolute Values of the Differences Between the Eight Layers in the Abundances of the Eight Types of Elements and Compounds

SO_4^{2-}	Cl^-	Br^+	I	Methane	S	C	N
0.1527	0.7608	0.3783	0.5378	0.1289	0.0199	0.6546	0.4473

Morishita-Horn [*Morisita*, 1959; *Horn*, 1966], *Ochiai* [1957], Sørensen [*Dice*, 1945; *Sørensen*, 1948], and *Whittaker* [1960] indices) and the absolute values of the differences between the layers in the abundances of the above eight types of elements and compounds. The correlation coefficients between the following pairs were greater than 0.3: the Anderberg index and Cl^- (0.3429); the Bray-Curtis index and Cl^-, I, C, and N (0.4463, 0.3170, 0.3237, and 0.3150); the Jaccard index and Cl^- (0.3431); the Morishita-Horn index and Cl^- (0.3023); the Ochiai index and Cl^- (0.3435); the Sørensen index and Cl^- (0.3434); and the Whittaker index and Cl^- (0.3434). There were no pairs between which the correlation coefficients were greater than 0.5. Although strong correlations were not found between the β diversity indices developed previously and the abundances of the studied elements, the elements with the highest correlations were those that were highly correlated with the β diversity index proposed in this study.

Therefore, we first examined the linear regression model with these four variables

$$\beta_{ii'} = b_0 + b_1|\Delta_{ii'}Cl^-| + b_2|\Delta_{ii'}I| + b_3|\Delta_{ii'}C| + b_4|\Delta_{ii'}N| + \varepsilon_{ii'}, \quad i = 1, \cdots, 7, i' = i+1, \cdots, 8 \tag{8}$$

as a model of the archaeal β diversity patterns. $\beta_{ii'}$ represents the β diversity between the archaeal communities in the ith and i'th layers, and $\Delta_{ii'}X$ denotes the difference in the abundances of X between the ith and i'th layers for each $X = Cl^-$, I, C, and N. $\varepsilon_{ii'}$ is an error term, and it is assumed that $\{\varepsilon_{ii'}\}$ has (i) no serial correlation and (ii) homoscedasticity and is (iii) the normal process, which will be tested below.

We estimated the model in equation (8) and the model without the intercept b_0 by the least-squares method. The result for the latter was relatively good and is given as follows (note that the equation without the intercept is more reasonable than the equation with the intercept as a β diversity pattern model):

$$\begin{aligned} \beta_{ii'} &= 1.3409|\Delta_{ii'}Cl^-| + 0.3304|\Delta_{ii'}I| + 26.0720|\Delta_{ii'}C| + 63.8382|\Delta_{ii'}N| \\ &\quad (0.0704) \qquad (0.0222) \qquad (0.3089) \qquad (0.7278) \\ F &= 43,0798(1.28\times10^{-10}),\ R^2 = 0.8778,\ \bar{R}^2 = 0.8574. \end{aligned}$$

For each $j = 1, \cdots, 4$, the value in the parentheses under the estimate of the regression coefficient b_j represents the p value for the t statistic for testing the hypothesis $b_j = 0$. F denotes the F statistic for testing $b_1 = b_2 = b_3 = b_4 = 0$ against the hypothesis that at least one b_j is different from 0, and the value in the parentheses to the right of F is the p value. R^2 and $\bar{R}^2$ represent the ordinary and adjusted coefficients of determination, respectively. Considering that this model is a model that has difference variables, not level variables, the values of R^2 and $\bar{R}^2$ were much higher than expected. Therefore, the above model had considerable explanatory power. All of the estimates of regression coefficients were positive, and thus, they satisfied the sign condition. From the p value for the F statistic, we could reject the hypothesis $b_1 = b_2 = b_3 = b_4 = 0$. However, the hypotheses $b_1 = 0$, $b_3 = 0$, and $b_4 = 0$ were not rejected because the p values for the t statistics for testing these hypotheses were large. Therefore, we could not insist that $|\Delta_{ii'}Cl^-|$, $|\Delta_{ii'}C|$, and $|\Delta_{ii'}N|$ were systematic factors for variations in $\beta_{ii'}$.

Therefore, we estimated several submodels of equation (8) and obtained the following result:

$$\begin{aligned} \beta_{ii'} &= 1.2040|\Delta_{ii'}Cl^-| + 0.3555|\Delta_{ii'}I| + 33.2933|\Delta_{ii'}C| \\ &\quad (0.0489) \qquad (0.0042) \qquad (0.0272) \\ F &= 59.4826(1.60\times10^{-11}),\ R^2 = 0.8771,\ \bar{R}^2 = 0.8624, \\ \mathrm{DW} &= 1.9089(0.3739),\ \mathrm{BG} = 0.0046(0.9459),\ \mathrm{BP} = 1.9996(0.3680), \\ \mathrm{GQ} &= 0.2884(0.9749),\ \mathrm{KS} = 0.1587(0.4358),\ \mathrm{RESET} = 2.3808(0.1149), \\ \mathrm{VIF}_1 &= 2.6274,\ \mathrm{VIF}_2 = 1.2613,\ \mathrm{VIF}_3 = 2.3200. \end{aligned} \tag{9}$$

All of the estimates of the regression coefficients were positive and satisfied the sign condition.

The hypotheses $b_j = 0$ for each $j = 1, 2, 3$ and $b_1 = b_2 = b_3 = 0$ were rejected because the p values for the t and F statistics were sufficiently small. In other words, all of the independent variables of the model in equation (9) could be statistically regarded as systematic factors for variations in the dependent variable $\beta_{ii'}$. Conversely, we tested whether the model was deficient in systematic factors for $\beta_{ii'}$ using Ramsey's RESET [*Ramsey*, 1974]. The test statistic and p value of this test were calculated as 2.3808 and 0.1149, respectively,

which did not suggest a deficiency in systematic factors for $\beta_{ii'}$. Therefore, we chose the three variables of $|\Delta_{ii'}Cl^-|$, $|\Delta_{ii'}I|$, and $|\Delta_{ii'}C|$ as systematic factors for the variation in $\beta_{ii'}$.

We next tested whether the error term $\epsilon_{ii'}$ satisfied the classical conditions under which the accuracy of least-squares estimates of regression coefficients is guaranteed. (i) DW and BG represent the test statistics of the Durbin-Watson test [*Durbin and Watson*, 1950, 1951] and Breusch-Godfrey test [*Breusch*, 1978; *Godfrey*, 1978] (tests of serial correlation), (ii) BP and GQ represent the test statistics of the Breusch-Pagan test [*Breusch and Pagan*, 1979] and Goldfeld-Quandt test [*Goldfeld and Quandt*, 1965] (tests of heteroscedasticity), and (iii) KS represents the test statistic of the Kolmogorov-Smirnov test [*Kolmogorov*, 1933; *Smirnov*, 1939] (a test of normality). The values provided in the parentheses to the right of these test statistics are *p* values. We also tested the higher-order serial correlation (up to the 7th order) of the model's error term using the Ljung-Box test [*Box and Pierce*, 1970; *Ljung and Box*, 1978]. For each $k = 1, \cdots, 7$, a test statistic LB (*k*) of the Ljung-Box test for *k*th order serial correlation and its *p* value are as follows: LB (1) = 0.0126 (0.9106), LB (2) = 1.5472 (0.4613), LB (3) =1.6708 (0.6434), LB (4) = 2.2556 (0.6889), LB (5) = 2.2584 (0.8124), LB (6) = 2.7268 (0.8423), LB (7) = 5.1480 (0.6419). From these values, we could proceed under the hypothesis that the error term in the above model satisfied the classical conditions for the least-squares estimators. Furthermore, there was no possibility of multicollinearity between the independent variables of the model because the variance inflation factor VIF_j for b_j was sufficiently small for each $j = 1, 2, 3$. Therefore, the least-squares estimates of the regression coefficients were reliable.

Figures 2a and 2b are the plots of $\beta_{ii'}$s and their estimates from the model in equation (9) and of the relative errors of the model in equation (9), respectively. From these figures, we can see that the estimates from the model explained variations in $\beta_{ii'}$ very well and that the relative errors of the model had a stable transition around 0. Furthermore, the ordinary and adjusted coefficients R^2 and $\bar{R}^2$ of determination were 0.8771 and 0.8624, respectively, and the above model explained 86–87% of the total variation in the archaeal β diversity between the layers below the seafloor. Therefore, the model presented in equation (9) for archaeal β diversity patterns had high performance.

5. Discussion

In the previous section, we found that the β diversity patterns between the archaeal communities in the eight layers below the seafloor off the east coast of the Shimokita Peninsula could be explained very well by the differences in the abundances of Cl^-, I, and C between the layers. Why was this result obtained? The β diversity between two communities is a type of distance between them. Therefore, there would be groups of archaea that are positively or negatively correlated with the abundances of these elements in the analyzed archaeal communities. In this section, we examine the basis for the observed explanatory power of the model of archaeal β diversity patterns by investigating the relationships between the composition of the archaeal communities and the abundances of the elements in the layers below the seafloor. Subsequently, we describe future challenges in modeling β diversity patterns along geochemical gradients and a further question raised by the results of this study.

We examined the origins of all archaeal ribosomal RNA gene sequences analyzed in the previous section using the SILVA database (http://www.arb-silva.de/). They were classified into 31 groups, as shown in Table 2. The information on the family and genus as well as the species could not be obtained for most sequences. Hereafter, we abbreviated these 31 archaeal groups by using the number in the leftmost column of Table 2. For example, Euryarchaeota (phylum) Methanobacteria (class) Methanobacteriales (order) Methanothermaceae (family) *Methanothermus* (genus) is Group 17. Table 3 provides the relative frequencies of the 31 archaeal groups in the eight layers. The relative frequencies that are greater than or equal to 5% are in bold. From this table, we see that Groups 5, 8, 22, and 29 of the 31 groups composed the majority of the sample. In the following paragraphs, we focus on these four groups because the relative frequencies of the sequences in the other groups were too small.

We calculated Pearson's correlation coefficients between the frequencies of the sequences of these four groups and the abundances of eight types of elements and compounds in the eight layers, as listed in the first paragraph of section 4.2. The absolute values of the correlation coefficients for the following four pairs were greater than 0.5: Group 5 and Cl^- (0.5942), Group 8 and Cl^- (−0.9629), Group 29 and Br^+ (−0.6326), and Group 29 and I (−0.6678). Figure 3a (left) provides the scatterplot of the pairs of the relative frequencies of the sequences in Group 5 in the eight layers and the abundances of Cl^- in the eight layers. The left panels

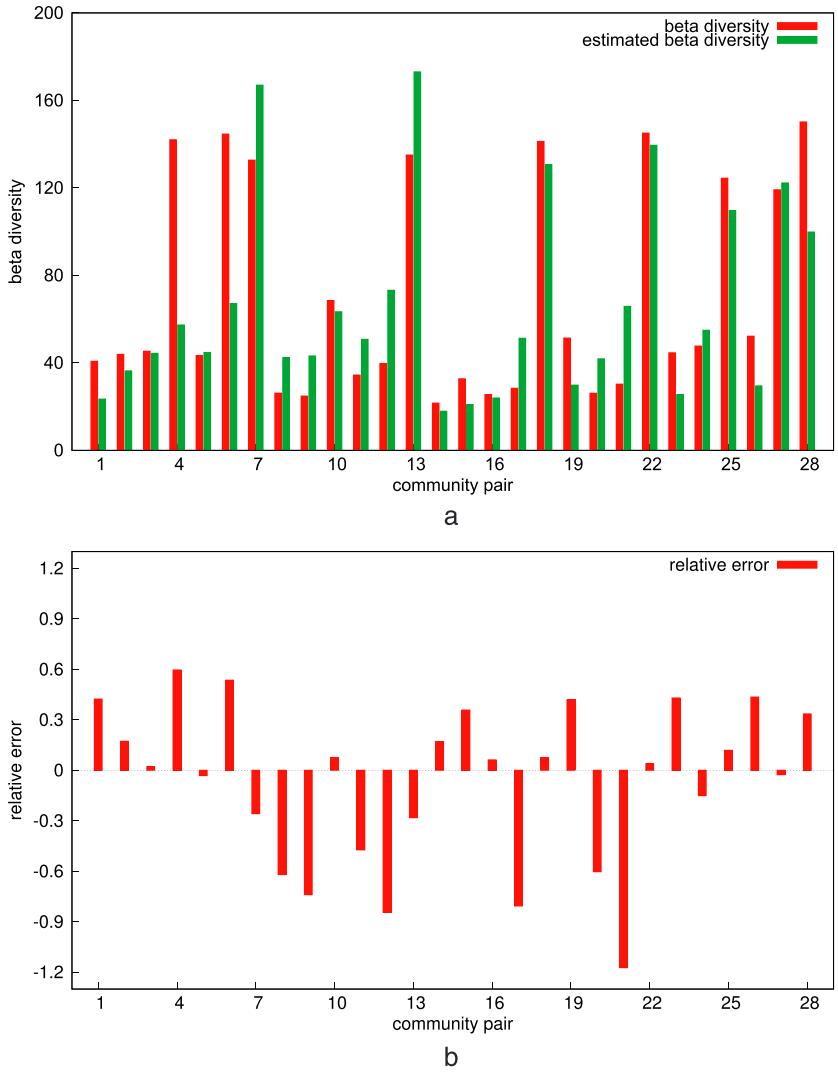

Figure 2. Plots (a) of the archaeal β diversities and their estimates from the model in equation (9) and (b) of the relative errors of the model in equation (9). Community pair 1 is the pair of archaeal communities in the first and second layers, pair 2 is the pair in the first and third layers, $\cdots$, and pair 28 is the pair in the seventh and eighth layers.

of Figures 3b, 3c, and 3d show the same scatterplots for the pairs of Group 8 and Cl^-, Group 29 and Br^+, and Group 29 and I, respectively. In these panels, the correlations between the frequencies of sequences and the abundances of elements are shown. Figure 3a (right) was prepared by plotting the pairs of the differences in the relative frequencies of the sequences in Group 5 between the layers and the differences in the abundances of Cl^- between the layers. The right panels of Figures 3b, 3c, and 3d were prepared for the pairs of Group 8 and Cl^-, Group 29 and Br^+, and Group 29 and I, respectively, in the same manner. In these right panels, the points in the first and third quadrants (shown in green) indicate a positive correlation, whereas the points in the second and fourth quadrants (shown in red) instead indicate a negative correlation. More than 75% of all 28 points were located in the first and third quadrants in Figure 3a (right) and in the second and fourth quadrants in the right panels of Figures 3b, 3c, and 3d. Therefore, these panels also support the presence of correlations between the frequencies of sequences and the abundances of elements. These correlations would underlie the high explanatory power of the model of archaeal β diversity patterns obtained in the previous section. Thus, the results in the previous section do not appear to be an artifact.

As described in section 1, the environmental variables such as (i) latitude, (ii) altitude, (iii) sea depth, (iv) temperature, (v) salinity, (vi) areas, and (vii) internal organs and skin of subjects have been used in the modeling of β diversity patterns in previous studies. For example, it is believed that a change in altitude leads

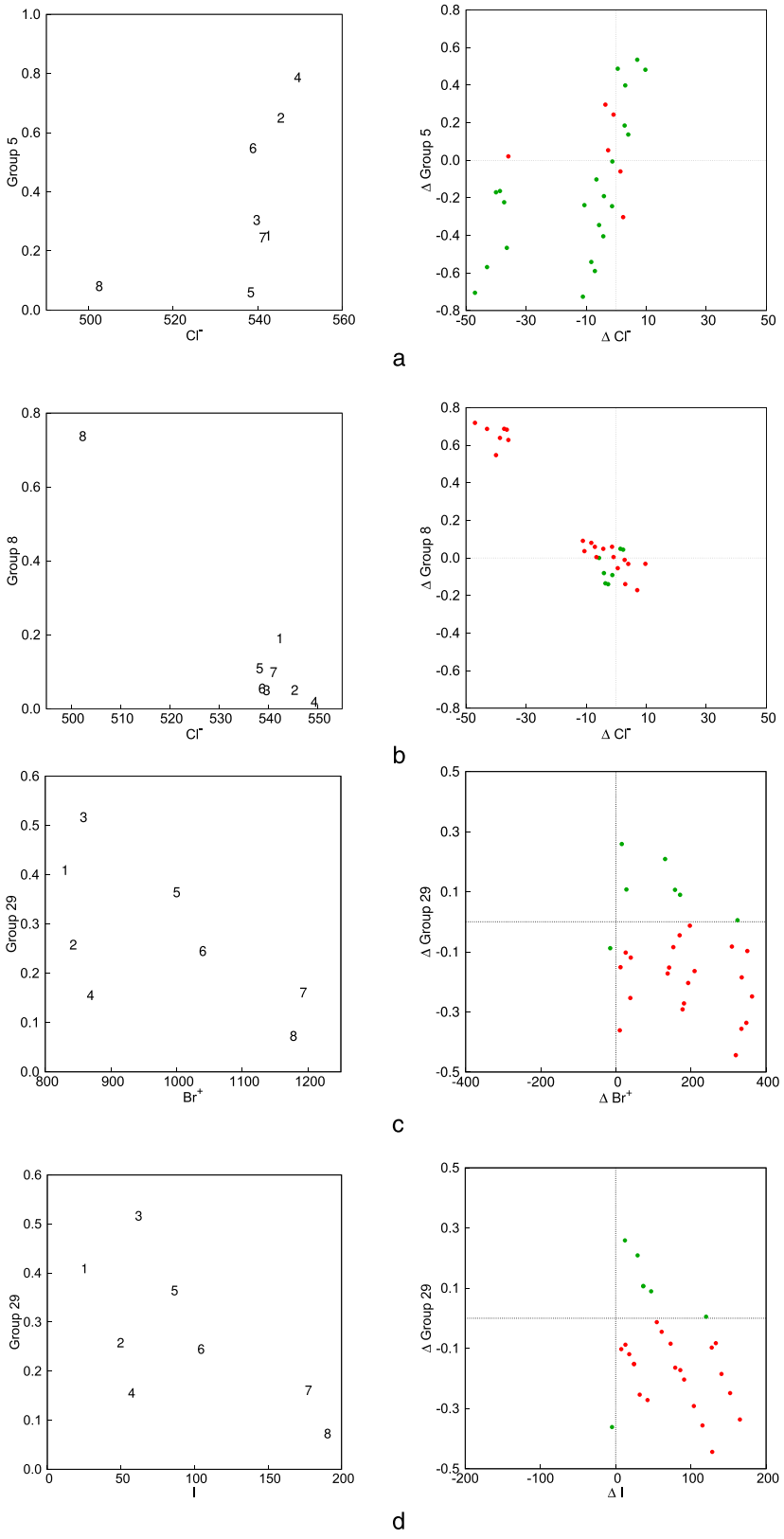

Figure 3. Scatterplots (a–d) of the pairs of the abundances of Cl^-, Br^+, and I in the eight layers and the relative frequencies of the sequences in Groups 5, 8, and 29 in the eight layers (left) and of the pairs of the differences in the abundances of Cl^-, Br^+, and I between the eight layers and the differences in the relative frequencies of the sequences in Groups 5, 8, and 29 between the eight layers (right).

Table 2. Classification of Archaeal Groups From Which the 16S Ribosomal RNA Gene Sequences in the Environmental Samples Originated

Group	Phylum	Class	Order	Family	Genus
1	Ancient Archaeal Group	unclassified	unclassified	unclassified	unclassified
2	Crenarchaeota	AK56	unclassified	unclassified	unclassified
3	Crenarchaeota	AK59	unclassified	unclassified	unclassified
4	Crenarchaeota	Group C3	unclassified	unclassified	unclassified
5	Crenarchaeota	Marine Benthic Group B	unclassified	unclassified	unclassified
6	Crenarchaeota	Marine Group I	Candidatus Nitrosopumilus	unclassified	unclassified
7	Crenarchaeota	Marine Group I	uncultured	unclassified	unclassified
8	Crenarchaeota	Miscellaneous Crenarchaeotic Group	unclassified	unclassified	unclassified
9	Crenarchaeota	Soil Crenarchaeotic Group	unclassified	unclassified	unclassified
10	Crenarchaeota	Terrestrial Hot Spring Group	unclassified	unclassified	unclassified
11	Crenarchaeota	Thermoprotei	Desulfurococcales	Pyrodictiaceae	*Pyrodictium*
12	Crenarchaeota	Z273FA48	unclassified	unclassified	unclassified
13	Euryarchaeota	Halobacteria	Halobacteriales	Deep Sea Euryarcheotic Group	unclassified
14	Euryarchaeota	Halobacteria	Halobacteriales	Halobacteriaceae	*Natronomonas*
15	Euryarchaeota	Halobacteria	Halobacteriales	Marine Hydrothermal Vent Group	unclassified
16	Euryarchaeota	Halobacteria	Halobacteriales	SM1K20	unclassified
17	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanothermaceae	*Methanothermus*
18	Euryarchaeota	Methanomicrobia	ANME-1	ANME-1a	unclassified
19	Euryarchaeota	Methanomicrobia	Methanocellales	BS-K-E9	unclassified
20	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methanosarcinaceae	*Methanococcoides*
21	Euryarchaeota	Methanomicrobia	Methanosarcinales	Methermicoccaceae	*Methermicoccus*
22	Euryarchaeota	Thermoplasmata	South African Goldmine Group	unclassified	unclassified
23	Euryarchaeota	Thermoplasmata	Thermoplasmatales	20c-4	unclassified
24	Euryarchaeota	Thermoplasmata	Thermoplasmatales	AMOS1A-4113-D04	unclassified
25	Euryarchaeota	Thermoplasmata	Thermoplasmatales	ANT06-05	unclassified
26	Euryarchaeota	Thermoplasmata	Thermoplasmatales	Amsterdam-1A-44	unclassified
27	Euryarchaeota	Thermoplasmata	Thermoplasmatales	CCA47	unclassified
28	Euryarchaeota	Thermoplasmata	Thermoplasmatales	MKCST-A3	unclassified
29	Euryarchaeota	Thermoplasmata	Thermoplasmatales	Marine Benthic Group D and DHVEG-1	unclassified
30	Euryarchaeota	Thermoplasmata	Thermoplasmatales	Marine Group III	unclassified
31	Euryarchaeota	Thermoplasmata	Thermoplasmatales	VC2.1 Arc6	unclassified

to changes in the temperature, pressure, and partial pressure of oxygen, and consequently, the community structure varies. Therefore, (i), (ii), (iii), (vi), and (vii) among these seven can be considered a composite of several variables. The point here is that using such composite variables as environmental variables in modeling β diversity patterns does not necessarily reveal key factors for determining β diversity patterns because they are composed of an unknown number of unspecified variables. Furthermore, the relations between β diversity and environmental variables are typically presented not by a numerical equation but in a graphical manner [see, for example, *Qian and Ricklefs*, 2007; *Miller et al.*, 2009]. However, using this graphical approach, it is difficult to increase the explanatory and predictive power of a model over the course of successive studies.

In this study, we adopted the abundances of elements, not composite variables, as environmental variables, and we presented the model of β diversity patterns in the form of a numerical equation with respect to these variables. Consequently, it became possible to take the research direction of refining the model through the examination of a function form and the selection of variables. With respect to the function form, according to the results in the previous section, the linear function works sufficiently well, but the reason for this positive result is not clear. Why the linear function can serve as an approximate model and whether a more appropriate function form can be found will be the topic of future research.

Another challenge is to provide a biogeochemical foundation for the selection of variables in the model. With regard to archaeal metabolism, methanogenesis and anaerobic methane oxidation in the Euryarchaeota and ammonia oxidation in the Crenarchaeota have been reported. See *Großkopf et al.* [1998], *Chouari et al.* [2005], and *Sakai et al.* [2007] for methanogenesis; *Brazelton et al.* [2006] and *Schleper* [2007] for anaerobic methane oxidation; and *Nunoura et al.* [2005], *Nicol and Schleper* [2006], *Schleper* [2007], *Teske and Sørensen* [2007], *de la Torre et al.* [2008], and *Hatzenpichler et al.* [2008] for ammonia oxidation. However, it has been revealed that uncultivated archaeal groups are dominant and that methanogens and anaerobic methane-oxidizing archaea are minorities under the seafloor [*Teske and Sørensen*, 2007]. Only a

Table 3. Relative Frequencies of the Sequences in the 31 Archaeal Groups in the Eight Layers

Layer	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
1	0.0297	0.0000	0.0000	**0.0503**	**0.2508**	0.0042	0.0014	**0.1899**
2	0.0093	0.0000	0.0000	0.0000	**0.6493**	0.0018	0.0027	**0.0502**
3	**0.0677**	0.0000	0.0000	0.0098	**0.3041**	0.0000	0.0017	0.0499
4	0.0117	0.0000	0.0000	0.0132	**0.7860**	0.0000	0.0000	0.0183
5	0.0108	0.0000	0.0000	**0.0573**	**0.0596**	0.0000	0.0000	**0.1092**
6	**0.0640**	0.0000	0.0000	0.0357	**0.5470**	0.0009	0.0000	**0.0546**
7	0.0059	0.0000	0.0000	0.0220	**0.2444**	0.0050	0.0000	**0.0984**
8	0.0086	0.0007	0.0094	0.0134	**0.0805**	0.0000	0.0000	**0.7371**
Layer	Group 9	Group 10	Group 11	Group 12	Group 13	Group 14	Group 15	Group 16
1	0.0028	0.0014	0.0021	0.0000	0.0021	0.0000	0.0354	0.0014
2	0.0009	0.0000	0.0000	0.0000	0.0000	0.0000	0.0167	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0115	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0066	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0263	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0489	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0031	0.0000	0.0015	0.0023	0.0023	0.0055	0.0000
Layer	Group 17	Group 18	Group 19	Group 20	Group 21	Group 22	Group 23	Group 24
1	0.0014	0.0000	0.0007	0.0007	0.0000	0.0021	0.0007	0.0000
2	0.0000	0.0000	0.0009	0.0000	0.0000	0.0093	0.0000	0.0009
3	0.0026	0.0008	0.0000	0.0000	0.0000	0.0276	0.0000	0.0000
4	0.0007	0.0022	0.0000	0.0000	0.0000	0.0007	0.0000	0.0014
5	0.0000	0.0023	0.0000	0.0000	0.0000	**0.3702**	0.0000	0.0000
6	0.0000	0.0009	0.0000	0.0000	0.0000	0.0018	0.0000	0.0000
7	0.0000	0.0050	0.0000	0.0000	0.0203	**0.4337**	0.0000	0.0000
8	0.0173	0.0047	0.0000	0.0000	0.0039	0.0363	0.0000	0.0000
Layer	Group 25	Group 26	Group 27	Group 28	Group 29	Group 30	Group 31	
1	0.0007	0.0014	0.0042	0.0035	**0.4089**	0.0014	0.0021	
2	0.0000	0.0000	0.0000	0.0000	**0.2576**	0.0000	0.0000	
3	0.0000	0.0053	0.0008	0.0008	**0.5165**	0.0000	0.0000	
4	0.0014	0.0000	0.0007	0.0014	**0.1551**	0.0000	0.0000	
5	0.0000	0.0000	0.0000	0.0000	**0.3640**	0.0000	0.0000	
6	0.0000	0.0000	0.0009	0.0000	**0.2448**	0.0000	0.0000	
7	0.0000	0.0000	0.0033	0.0008	**0.1604**	0.0000	0.0000	
8	0.0000	0.0000	0.0000	0.0000	**0.0726**	0.0000	0.0000	

minor portion of the genetic information of uncultivated diverse archaeal groups has been obtained, with the exception of ribosomal small-subunit RNAs, and details of their metabolism have not been revealed. Therefore, it is difficult to provide variable selection in the model of the archaeal β diversity patterns with a biogeochemical foundation based on the present knowledge of archaeal metabolism.

However, results that support the selection of variables in our model were reported in a previous study. *Lozupone and Knight* [2007] clustered the environmental samples of 21,752 bacterial 16S ribosomal RNA sequences compiled from 111 studies of various environments by the phylogenetic lineages that they contain by applying principal coordinate analysis [*Gower*, 1966] and hierarchical clustering [*Sokal and Michener*, 1958] to a matrix of UniFrac [*Lozupone and Knight*, 2005] distances. They reported that a key factor for determining the presence or absence of bacterial lineages among environments was salinity rather than temperature, pH, or other factors represented in their samples, although determinations of salinity in their study are not direct measurements of salt concentration but instead are qualitative and based on the habitat descriptions. This finding supports the inclusion of Cl^- in the environmental variables of the model of β diversity patterns because Cl^- is a proxy of salinity. The correlation between Br^+ and Group 29 indicated by Figure 3c and the correlation coefficient in the third paragraph of section 5 also is of interest. Br and the variables Cl and I in the archaeal β diversity pattern model are all halogens. In previous studies on archaeal metabolism, little attention has been paid to these halogens, which exist in bulk in seawater and the crust, compared to the attention that carbon, nitrogen, and sulfur have received. A negative correlation between the frequency of an archaeal group and the abundance of an element or compound might arise if the group avoids environments where the element or compound is abundant. A positive correlation might

be observed if the group produces energy from the element or compound (or if the group produces the element or compound as a metabolic product, although this possibility is less likely). Novel findings are awaited with respect to the metabolism of the above archaeal groups.

In this study, we defined the β diversity index between two biological communities according to the extent to which the total sequences of individuals in the two communities varied. We constructed the estimator of the β diversity index based on sequence data and, using the theory of random strings developed by *Koyano and Kishino* [2010], demonstrated that the proposed estimator converges to the β diversity index between populations with probability of 1 as the number of collected sequences increases. Furthermore, we applied the developed method to environmental samples of 16S ribosomal RNA gene sequences collected from archaeal communities in the layers below the seafloor off the east coast of the Shimokita Peninsula, and we described the estimated archaeal β diversity patterns using a quantitative equation incorporating variables of the differences between layers in the abundances of several elements. It is reasonable to predict that archaeal communities in layers vary more the farther apart the layers are because layers are deposited in chronological order. However, as shown in section 4.2, a more important determinant of the β diversity patterns of the archaeal communities in the layers below the seafloor off the east coast of the Shimokita Peninsula is the differences in the abundances of Cl^-, I, and C between the layers rather than the order of the layers. This finding suggests that the elements in the surrounding environment strongly influence the evolution of archaeal communities.

In general, previous studies of β diversity have measured β diversity between communities in different areas at a fixed point in time and have attempted to identify the determinants of the measured β diversity. The present study also models archaeal β diversity patterns within this time-fixed, varied-area framework. In contrast to this framework, by fixing area rather than time, we can consider the problem of developing an equation that describes how a microbial community interacts with the environment in an area and how both the community and environment change over time. In the long history of life on earth, while microorganisms have evolved to adapt to environments, they have also altered environments by forming metabolic systems. Therefore, such an equation would be fundamental in biogeoscience if it can be obtained. However, it is impossible to collect environmental samples of 16S ribosomal RNA gene sequences from microbial communities and to measure the abundances of various elements and compounds in an environment over the long term. Instead, can we approach this problem in an alternative way using environmental samples of 16S ribosomal RNA gene sequences collected from microbial communities and data that represent material abundance in different layers, as analyzed in this study? It is the next methodological challenge to extend the method and approach used in this study toward constructing a model that describes the variation of an environment and the evolution of a microbial community with time through their interaction using the information on the age of layers and incorporating corrections to consider material flux and the movement of microorganisms by flux between layers.

Appendix A: Summary of Probability Theory for Strings and a Proof of the Strong Consistency of the β Diversity Estimator

A1. Random Strings

In this subsection, we describe definitions for several concepts in probability theory on a set of strings used in section 3. See the online supplemental material of *Koyano and Kishino* [2010] for details. In the following paragraphs, we refer to a set of a finite number of letters

$$A = \{a_1, \cdots, a_{c-1}\}$$

as the alphabet. For example, $A = \{a, c, g, t\}$ is an alphabet for gene sequences. Let us denote the empty letter by e. We set $\bar{A} = A \cup \{e\}$ and call $\bar{A}$ the extended alphabet. We set $A^k = \{(x_1, \cdots, x_k) : x_1, \cdots, x_k \in A\}$ for any $k \in \mathbb{Z}^+$. $\bar{A}^k$ is defined in a similar manner. Let $(\Omega, \mathfrak{F}, P)$ be a probability space. We denote the power set of a set X by 2^X. We refer to an $\bar{A}$-valued random variable on Ω as a random letter and denote the set of all random letters by $\mathcal{M}(\Omega, \bar{A})$, that is,

$$\mathcal{M}(\Omega, \bar{A}) = \{\alpha : \Omega \to \bar{A} : \alpha^{-1}(B) \in \mathfrak{F}, B \in 2^{\bar{A}}\}.$$

For the mapping $\epsilon : \Omega \to \bar{A}$ that is defined as $\epsilon(\omega) = e$ for any $\omega \in \Omega$, we have $\epsilon \in \mathcal{M}(\Omega, \bar{A})$. In common usage in computer science, a string on the alphabet $A = \{a_1, \cdots, a_{c-1}\}$ is a finite sequence of elements of A. However, in this study, we define a string as follows, although both definitions are essentially identical.

Definition 3: A sequence $s = \{x_i \in \bar{A} : i \in \mathbb{Z}^+\}$ of elements of $\bar{A}$ is a string on A if it satisfies the following conditions:

$$\text{(i) there exists } k \in \mathbb{Z}^+ \text{ such that } x_k = e, \text{ and (ii) } x_\ell = e \text{ implies } x_{\ell+1} = e.$$

In other words, we define a string on A as a finite sequence of elements of A to which the infinite sequence $(e, \cdots)$ of the empty letter is appended. In the following definition (Definition 4), by naturally extending the above definition of a string, we define a random string in a manner in which it can realize strings of varying lengths. We denote the set of all strings on A by A^*. A function $|\cdot| : A^* \to \mathbb{N}$ is defined as

$$|s| = \min\{j \in \mathbb{Z}^+ : x_j = e\} - 1,\ s = \{x_i : i \in \mathbb{Z}^+\}$$

and called the length on A^*.

Definition 4: A sequence of random letters $\sigma = \{\alpha_i \in \mathcal{M}(\Omega, \bar{A}) : i \in \mathbb{Z}^+\}$ is a random string if it satisfies the following conditions:

(i) for any $\omega \in \Omega$ there exists $k \in \mathbb{Z}^+$ such that $\alpha_k(\omega) = e$, and

(ii) $\alpha_\ell(\omega) = e$ for $\omega \in \Omega$ implies $\alpha_{\ell+1}(\omega) = e$.

We denote the set of all random strings by $\mathcal{M}(\Omega, A^*)$. A function $|\cdot| : \mathcal{M}(\Omega, A^*) \to \mathbb{N}$ is defined as

$$|\sigma| = \min\{j \in \mathbb{Z}^+ : \alpha_j = \epsilon\} - 1,\ \sigma = \{\alpha_i : i \in \mathbb{Z}^+\}$$

and called the length on $\mathcal{M}(\Omega, A^*)$. The random string that was defined in Definition 4 can be regarded as a special case of a discrete stochastic process. Therefore, the finite-dimensional distribution of a random string and the independence of random strings are defined as follows.

Definition 5: Let $\sigma = \{\alpha_i : i \in \mathbb{Z}^+\} \in \mathcal{M}(\Omega, A^*)$. A set function $\boldsymbol{Q}_{\sigma;i_1,\cdots,i_k} : 2^{\bar{A}^k} \to [0, 1]$ is defined as

$$\boldsymbol{Q}_{\sigma;i_1,\cdots,i_k}(B) = P\left(\left\{\omega \in \Omega : (\alpha_{i_1}(\omega), \cdots, \alpha_{i_k}(\omega)) \in B\right\}\right)$$

for each $k \in \mathbb{Z}^+$ and $i_1, \cdots, i_k \in \mathbb{Z}^+$ that satisfies $i_1 < \cdots < i_k$. $\boldsymbol{Q}_{\sigma;i_1,\cdots,i_k}$ is a probability measure on $2^{\bar{A}^k}$ and is called the finite-dimensional distribution of σ at sites $i_1, \cdots, i_k$. A function $\boldsymbol{q}_{\sigma;i_1,\cdots,i_k} : \bar{A}^k \to [0, 1]$ is defined as

$$\boldsymbol{q}_{\sigma;i_1,\cdots,i_k}(x_1, \cdots, x_k) = \boldsymbol{Q}_{\sigma;i_1,\cdots,i_k}\left(\{(x_1, \cdots, x_k)\}\right)$$

and called the probability function of $\boldsymbol{Q}_{\sigma;i_1,\cdots,i_k}$.

Definition 6: (1) Finite case. $\sigma_1 = \{\alpha_{1j} : j \in \mathbb{Z}^+\}, \cdots, \sigma_n = \{\alpha_{nj} : j \in \mathbb{Z}^+\} \in \mathcal{M}(\Omega, A^*)$ are independent if $(\alpha_{1j} : j \in I_1), \cdots, (\alpha_{nj} : j \in I_n)$ are independent for any nonempty finite set $I_1, \cdots, I_n \subset \mathbb{Z}^+$. (2) Countably infinite case. $\{\sigma_i : i \in \mathbb{Z}^+\} \subset \mathcal{M}(\Omega, A^*)$ are independent if $\sigma_{i_1}, \cdots, \sigma_{i_k}$ are independent for any $k \in \mathbb{Z}^+$ and $i_1, \cdots, i_k \in \mathbb{Z}^+$.

A2. Proof of Proposition 1

In this subsection, the proof of Proposition 1 described in section 3.2 is provided.

We denote the number of elements of a countable set X by $\sharp X$. Setting $W_k = \{s \in A^* : |s| = k\}$ for any $k \in \mathbb{N}$, we have $\sharp W_k = (c-1)^k$. Therefore, the number of strings whose length is less than or equal to ℓ can be represented as $\sum_{k=0}^{\ell}(c-1)^k$. Thus, noting that $\sum_{k=0}^{\ell}(c-1)^k < \infty$ holds for $\ell < \infty$ and the definition of a string (Definition 3), we have $\sharp A^* < \infty$, and thus, there exists $r \in \mathbb{Z}^+$ such that we can write $A^* = \{u_1, \cdots, u_r\}$. We first consider the population that has the probability function $\boldsymbol{p}$. We define an r-dimensional random vector $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ir})$ by setting

$$X_{ij} = 1 \text{ and } X_{ij'} = 0 \text{ for } j' \in \{1, \cdots, r\} - \{j\}$$

if u_j is observed in the ith observation from this population for each $i \in \{1, \cdots, m\}$. $\boldsymbol{X}_i$ has a multinomial distribution with the number of trials 1 and the success probabilities $\boldsymbol{p}(u_1), \cdots, \boldsymbol{p}(u_r)$, and therefore, the

expectation vector of $\boldsymbol{X}_i$ is given by $(\boldsymbol{p}(u_1), \cdots, \boldsymbol{p}(u_r))$. Because $\sigma_1, \cdots, \sigma_m$ are independent, $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ are also independent. Thus, applying the strong law of large numbers in $\mathbb{R}^r$, we have

$$\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{X}_i \xrightarrow{\text{a.s.}} (\boldsymbol{p}(u_1), \cdots, \boldsymbol{p}(u_r))\ (m \longrightarrow \infty),$$

where $\xrightarrow{\text{a.s.}}$ represents almost sure convergence. Thus, noting $(1/m)\sum_{i=1}^{m}\boldsymbol{X}_i = (\hat{\boldsymbol{p}}_{S^{(m)}}(u_1), \cdots, \hat{\boldsymbol{p}}_{S^{(m)}}(u_r))$ gives

$$\hat{\boldsymbol{p}}_{S^{(m)}}(u_j) \xrightarrow{\text{a.s.}} \boldsymbol{p}(u_j)\ (m \longrightarrow \infty) \tag{A1}$$

for each $j \in \{1, \cdots, r\}$. For the other population that has the probability function $\boldsymbol{q}$, we obtain

$$\hat{\boldsymbol{q}}_{T^{(n)}}(u_j) \xrightarrow{\text{a.s.}} \boldsymbol{q}(u_j)\ (n \longrightarrow \infty) \tag{A2}$$

for each $j \in \{1, \cdots, r\}$ in the same manner. Noting equations (A1) and (A2) and applying the continuous mapping theorem [*Mann and Wald*, 1943] leads to

$$|\hat{\boldsymbol{p}}_{S^{(m)}}(u_j) - \hat{\boldsymbol{q}}_{T^{(n)}}(u_j)| \xrightarrow{\text{a.s.}} |\boldsymbol{p}(u_j) - \boldsymbol{q}(u_j)|\ (m, n \longrightarrow \infty). \tag{A3}$$

We have $\sharp D_{\boldsymbol{p}} < \infty$ for the support $D_{\boldsymbol{p}}$ of $\boldsymbol{p}$ because $D_{\boldsymbol{p}} \subset A^*$, and therefore, we write $D_{\boldsymbol{p}} = \{u'_1, \cdots, u'_{r'}\}$. Combining equation (A1) and $\boldsymbol{p}(u'_1), \cdots, \boldsymbol{p}(u'_{r'}) > 0$, we see that for any $j \in \{1, \cdots, r'\}$, there exists $m_j \in \mathbb{Z}^+$ such that if $m \geq m_j$,

$$X_{ij} \geq 1 \quad \text{a.s.}$$

holds for at least one $i \in \{1, \cdots, m\}$, where a.s. represents that a statement in front of it holds with probability of 1. Thus, setting $m^* = \max\{m_1, \cdots, m_{r'}\}$, we have $S^{(m)} = D_{\boldsymbol{p}}$ a.s. for any $m \geq m^*$. We find that if we choose a sufficiently large $n^* \in \mathbb{Z}^+$, $T^{(n)} = D_{\boldsymbol{q}}$ a.s. holds for any $n \geq n^*$ in the same manner. Thus, if $m \geq m^*$ and $n \geq n^*$ for such m^* and n^*,

$$\hat{\delta}_{S^{(m)}, T^{(n)}}(s) = \delta_{\boldsymbol{p},\boldsymbol{q}}(s) \quad \text{a.s.} \tag{A4}$$

holds. Noting equations (6), (7), (A3), and (A4) and applying the continuous mapping theorem, we obtain

$$\widehat{d}_\beta(\boldsymbol{p}, \boldsymbol{q}) \xrightarrow{\text{a.s.}} d_\beta(\boldsymbol{p}, \boldsymbol{q})\ (m, n \longrightarrow \infty).$$

Thus, the strong consistency of $\widehat{d}_\beta(\boldsymbol{p}, \boldsymbol{q})$ has been demonstrated.

Acknowledgments
We are grateful to anonymous referees for valuable comments on the manuscript. This work was supported in part by Grant-in-Aid for Challenging Exploratory Research from the Japan Society for the Promotion of Science (26610037). Computational resources were provided by Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

Anderberg, M. R. (1973), *Cluster Analysis for Applications*, Academic Press, New York.
Anderson, M. J., K. E. Ellingsen, and B. H. McArdle (2006), Multivariate dispersion as a measure of beta diversity, *Ecol. Lett.*, *9*, 683–693.
Aoike, K. (2007), *CK06-06 D/V Chikyu Shakedown Cruise Offshore, Shimokita Laboratory Operation Report*, Science and Planning Department, Center for Deep Earth Exploration and Japan Agency for Mrine–Earth Science and Technology, Tokyo.
Box, G. E. P., and D. A. Pierce (1970), Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *J. Am. Stat. Assoc.*, *65*, 1509–1526.
Bray, J. R., and J. T. Curtis (1957), An ordination of the upland forest communities of southern Wisconsin, *Ecol. Monogr.*, *27*, 325–349.
Brazelton, W. J., M. O. Schrenk, D. S. Kelley, and J. A. Baross (2006), Methane- and sulfur-metabolizing microbial communities dominate the Lost City hydrothermal field ecosystem, *Appl. Environ. Microbiol.*, *72*, 6257–6270.
Brehm, G., J. Homeier, and K. Fiedler (2003), Beta diversity of geometrid moths (Lepidoptera: Geometridae) in an Andean montane rainforest, *Diversity Distrib.*, *9*, 351–366.
Breusch, T. S. (1978), Testing for autocorrelation in dynamic linear models, *Aust. Econ. Pap.*, *17*, 334–355.
Breusch, T. S., and A. R. Pagan (1979), Simple test for heteroscedasticity and random coefficient variation, *Econometrica*, *47*, 1287–1294.
Chao, A., R. L. Chazdon, R. K. Colwell, and T.-J. Shen (2005), A new statistical approach for assessing similarity of species composition with incidence and abundance data, *Ecol. Lett.*, *8*, 148–159.
Chouari, R., D. Le Paslier, P. Daegelen, P. Ginestet, J. Weissenbach, and A. Sghir (2005), Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester, *Environ. Microbiol.*, *7*, 1104–1115.
Cody, M. L. (1975), Towards a theory of continental species diversities: Bird distributions over mediterranean habitat gradients, in *Ecology and Evolution of Communities*, edited by M. L. Cody, and J. M. Diamond, pp. 214–257, Belknap Press, Harvard.
Cody, M. L. (1993), Bird diversity components within and between habitats in Australia, in *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*, edited by R. E. Ricklefs, and D. Schluter, pp. 147–158, Univ. of Chicago Press, Chicago.
Colwell, R. K., and J. A. Coddington (1994), Estimating terrestrial biodiversity through extrapolation, *Phil. Trans. R. Soc. B*, *345*, 101–118.
Condit, R., et al. (2002), Beta-diversity in tropical forest trees, *Science*, *295*, 666–669.

Czekanowski, J. (1909), Zur differential Diagnose der Neandertalgruppe, *Korrespbl. dt. Ges. Anthrop.*, *40*, 44–47.
de la Torre, J. R., C. B. Walker, A. E. Ingalls, M. Könneke, and D. A. Stahl (2008), Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol, *Environ. Microbiol.*, *10*, 810–818.
D'Hondt, S., et al. (2004), Distributions of microbial activities in deep subseafloor sediments, *Science*, *306*, 2216–2221.
Dice, L. R. (1945), Measures of the amount of ecologic association between species, *Ecology*, *26*, 297–302.
Durbin, J., and G. S. Watson (1950), Testing for serial correlation in least squares regression. I, *Biometrika*, *37*, 409–428.
Durbin, J., and G. S. Watson (1951), Testing for serial correlation in least squares regression. II, *Biometrika*, *38*, 159–177.
Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman (2005), Diversity of the human intestinal microbial flora, *Science*, *308*, 1635–1638.
Elderfield, H., C. G. Wheat, M. J. Mottl, C. Monnin, and B. Spiro (1999), Fluid and geochemical transport through oceanic crust: A transect across the eastern flank of the Juan de Fuca Ridge, *Earth Planet. Sci. Lett.*, *172*, 151–165.
Gao, Z., C. Tseng, Z. Pei, and M. J. Blaser (2007), Molecular analysis of human forearm superficial skin bacterial biota, *Proc. Natl. Acad. Sci. U. S. A.*, *104*, 2927–2932.
Gaston, K. J., A. S. L. Rodrigues, B. J. van Rensburg, P. Koleff, and S. L. Chown (2001), Complementary representation and zones of ecological transition, *Ecol. Lett.*, *4*, 4–9.
Godfrey, L. G. (1978), Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables, *Econometrica*, *46*, 1303–1310.
Goldfeld, S. M., and R. E. Quandt (1965), Some tests for homoscedasticity, *J. Am. Stat. Assoc.*, *60*, 539–547.
Gower, J. C. (1966), Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, *53*, 325–338.
Gower, J. C. (1971), A general coefficient of similarity and some of its properties, *Biometrics*, *27*, 857–871.
Großkopf, R., P. H. Janssen, and W. Liesack (1998), Diversity and structure of the methanogenic community in anoxic rice paddy soil microcosms as examined by cultivation and direct 16S rRNA gene sequence retrieval, *Appl. Environ. Microbiol.*, *64*, 960–969.
Harrison, S., S. J. Ross, and J. H. Lawton (1992), Beta diversity on geographic gradients in Britain, *J. Anim. Ecol.*, *61*, 151–158.
Harte, J., and A. P. Kinzig (1997), On the implications of species-area relationships for endemism, spatial turnover, and food web patterns, *Oikos*, *80*, 417–427.
Hatzenpichler, R., E. V. Lebedeva, E. Spieck, K. Stoecker, A. Richter, H. Daims, and M. Wagner (2008), A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring, *Proc. Natl. Acad. Sci. USA*, *105*, 2134–2139.
Horn, H. S. (1966), Measurement of "overlap" in comparative ecological studies, *Am. Nat.*, *100*, 419–424.
Izsak, C., and A. R. G. Price (2001), Measuring β-diversity using a taxonomic similarity index, and its relation to spatial scale, *Mar. Ecol. Prog. Ser.*, *215*, 69–77.
Jaccard, P. (1900), Contribution au problème de l'immigration post-glaciare de la flore alpine, *Bull. Soc. Vaudoise Sci. Nat.*, *36*, 87–130.
Koleff, P., K. J. Gaston, and J. J. Lennon (2003), Measuring beta diversity for presence-absence data, *J. Anim. Ecol.*, *72*, 367–382.
Kolmogorov, A. N. (1933), Sulla determinazione empirica di una legge di distributione, *Giornale dell'flstituto Italiano degli Attuari*, *4*, 1–11.
Koyano, H., and H. Kishino (2010), Quantifying biodiversity and asymptotics for a sequence of random strings, *Phys. Rev. E*, *81*, 061912.
Lande, R. (1996), Statistics and partitioning of species diversity, and similarity among multiple communities, *Oikos*, *76*, 5–13.
Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston (2001), The geographical structure of British bird distributions: Diversity, spatial turnover and scale, *J. Anim. Ecol.*, *70*, 966–979.
Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Sov. Phys. Dokl.*, *10*, 707–710.
Lipp, J. S., Y. Morono, F. Inagaki, and K. U. Hinrichs (2008), Significant contribution of Archaea to extant biomass in marine subsurface sediments, *Nature*, *454*, 991–994.
Ljung, G. M., and G. E. P. Box (1978), On a measure of lack of fit in time series models, *Biometrika*, *65*, 297–303.
Lozupone, C., and R. Knight (2005), UniFrac: A new phylogenetic method for comparing microbial communities, *Appl. Environ. Microbiol.*, *71*, 8228–8235.
Lozupone, C. A., and R. Knight (2007), Global patterns in bacterial diversity, *Proc. Natl. Acad. Sci. USA*, *104*, 11,436–11,440.
Lozupone, C. A., and R. Knight (2008), Species divergence and the measurement of microbial diversity, *FEMS Microbiol. Rev.*, *32*, 557–578.
Lozupone, C. A., M. Hamady, S. T. Kelley, and R. Knight (2007), Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities, *Appl. Environ. Microbiol.*, *73*, 1576–1585.
Magurran, A. E. (2004), *Measuring Biological Diversity*, Blackwell, Oxford, U. K.
Manheim, F. T., E. G. Brooks, and W. J. Winters (1994), Description of hydraulic sediment squeezer, *Tech. Rep. 94-0584*, U. S. Geological Survey Open-File Report, Woods Hole, Mass.
Mann, H. B., and A. Wald (1943), On stochastic limit and order relationships, *Ann. Math. Stat.*, *14*, 217–226.
Miller, S. R., A. L. Strong, K. L. Jones, and M. C. Ungerer (2009), Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park, *Appl. Environ. Microbiol.*, *75*, 4565–4572.
Morisita, M. (1959), Measuring of interspecific association and similarity between communities, *Mem. Fac. Sci. Kyushu Univ. E*, *3*, 65–80.
Motyka, J. (1947), *O zadaniach i metodach badan geobotanicznych. sur les buts et les méthodes des recherches géobotaniques*, Annales Universitatis Mariae Curie-Sklodowska, Sectio C, Supplementum I, Nakladem Uniwersytetu Marii Curie-Sklodowskiej, Lublin, Poland.
Mourelle, C., and E. Ezcurra (1997), Differentiation diversity of Argentine cacti and its relationship to environmental factors, *J. Veg. Sci.*, *8*, 547–558.
Nakanishi, M., K. Tamaki, and K. Kobayashi (1989), Mesozoic magnetic anomaly lineations and seafloor spreading history of the northwestern Pacific, *J. Geophys. Res.*, *94*, 15,437–15,462.
Nakanishi, M., K. Tamaki, and K. Kobayashi (1992), A new Mesozoic isochron chart of the northwestern Pacific Ocean: Paleomagnetic and tectonic implications, *Geophys. Res. Lett.*, *19*, 693–696.
Nicol, G. W., and C. Schleper (2006), Ammonia-oxidising Crenarchaeota: Important players in the nitrogen cycle? *Trends Microbiol.*, *14*, 207–212.
Nunoura, T., et al. (2005), Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments, *Environ. Microbiol.*, *7*, 1967–1984.
Nunoura, T., et al. (2011), Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group, *Nucleic Acids Res.*, *39*, 3204–3223.
Ochiai, A. (1957), Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Jpn. Soc. Sci. Fish*, *22*, 526–530.
Odum, E. P. (1950), Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion, *Ecology*, *31*, 587–605.

Parkes, R. J., B. A. Cragg, S. J. Bale, J. M. Getliff, K. Goodman, P. A. Rochelle, J. C. Fry, A. J. Weightman, and S. M. Harvey (1994), Deep bacterial biosphere in Pacific Ocean sediments, *Nature*, *371*, 410–413.

Pavoine, S., A. B. Dufour, and D. Chessel (2004), From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis, *J. Theor. Biol.*, *228*, 523–537.

Plotkin, J. B., and H. C. Muller-Landau (2002), Sampling the species composition of a landscape, *Ecology*, *83*, 3344–3356.

Qian, H., and R. E. Ricklefs (2007), A latitudinal gradient in large-scale beta diversity for vascular plants in North America, *Ecol. Lett.*, *10*, 737–744.

Ramsey, J. B. (1974), Classical model selection through specification error tests, in *Frontiers in Econometrics*, edited by P. Zarembka, pp. 13–47, Academic Press, New York.

Rao, C. R. (1982), Diversity and dissimilarity coefficients: A unified approach, *Theor. Pop. Biol.*, *21*, 24–43.

Renkonen, O. (1938), Statistisch-ökologische untersuchungen über die terrestrische käferwelt der finnischen bruchmoore, *Ann. Zool. Soc. Zool. –Bot. Fennicae Vanamo*, *6*, 1–123.

Rodriguez, P., and H. T. Arita (2004), Beta diversity and latitude in North American mammals: Testing the hypothesis of covariation, *Ecography*, *27*, 547–556.

Routledge, R. D (1977), On Whittaker's components of diversity, *Ecology*, *58*, 1120–1127.

Ruggiero, A., J. H. Lawton, and T. M. Blackburn (1998), The geographic ranges of mammalian species in south america: Spatial patterns in environmental resistance and anisotropy, *J. Biogeogr.*, *25*, 1093–1103.

Sakai, S., H. Imachi, Y. Sekiguchi, A. Ohashi, H. Harada, and Y. Kamagata (2007), Isolation of key methanogens for global methane emission from rice paddy fields: A novel isolate affiliated with the clone cluster rice cluster I, *Appl. Environ. Microbiol.*, *73*, 4326–4331.

Santoro, A. E., A. B. Boehm, and C. A. Francis (2006), Denitrifier community composition along a nitrate and salinity gradient in a coastal aquifer, *Appl. Environ. Microbiol.*, *72*, 2102–2109.

Schleper, C. (2007), Diversity of uncultivated archaea: Perspectives from microbial ecology and metagenomics, in *Archaea Evolution, Physiology, and Molecular Biology*, edited by R. Garrett, and H.-P. Klenk, Blackwell, Malden.

Shiryaev, A. N. (1996), *Probability*, Springer, New York.

Smirnov, H. (1939), Sur les Écarts de la courbe de distribution empirique, *Recl. Math.*, *6*, 3–26.

Smith, W., A. R. Solow, and P. E. Preston (1996), An estimator of species overlap using a modified beta-binomial model, *Biometrics*, *52*, 1472–1477.

Sokal, R. R., and C. D. Michener (1958), A statistical method of evaluating systematic relationships, *Univ. Kansas Sci. Bull.*, *28*, 1409–1438.

Sørensen, T. A. (1948), A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, *5*, 1–34.

Southwood, T. R. E., and P. A. Henderson (2000), *Ecological Methods*, Wiley-Blackwell, Oxford U. K.

Taira, A., D. Curewitz, T. Yohroh, A. Hashimoto, A. Ibusuki, T. Maruyama, T. Okano, S. Sasakawa, and H. Tanaka (2005), Shimokita Area Site Survey: Northern Japan Trench Seismic Survey, Offshore Northern Honshu, Japan, *Tech. Rep. 2*, CDEX/JAMSTEC, Yokohama, Japan.

Teske, A., and K. B. Sørensen (2007), Uncultured archaea in deep marine subsurface sediments: Have we caught them all? *ISME J.*, *2*, 3–18.

Tomaru, H., U. Fehn, Z. Lu, R. Takeuchi, F. Inagaki, H. Imachi, R. Kotani, R. Matsumoto, and K. Aoike (2009), Dating of dissolved iodine in pore waters from the gas hydrate occurrence offshore Shimokita Peninsula, Japan: ^{129}I results from the D/V Chikyu shakedown cruise, *Resour. Geol.*, *59*, 359–373.

Torgerson, W. S. (1952), Multidimensional scaling: I. Theory and method, *Psychometrika*, *17*, 401–419.

von Huene, R., and R. Culotta (1989), Tectonic erosion at the front of the Japan Trench convergent margin, *Tectonophysics*, *160*, 75–90.

Walsh, D. A., R. T. Papke, and W. F. Doolittle (2005), Archaeal diversity along a soil salinity gradient prone to disturbance, *Environ. Microbiol.*, *7*, 1655–1666.

Weiher, E., and C. W. Boylen (1994), Patterns and prediction of α and β diversity of aquatic plants in Adirondack (New York) lakes, *Can. J. Bot.*, *72*, 1797–1804.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe (1998), Prokaryotes: The unseen majority, *Proc. Natl. Acad. Sci. USA*, *95*, 6578–6583.

Whittaker, R. H. (1960), Vegetation of the Siskiyou Mountains, Oregon and California, *Ecol. Monogr.*, *30*, 407–407.

Whittaker, R. H. (1972), Evolution and measurement of species diversity, *Taxon*, *21*, 213–251.

Williams, P. H. (1996), Mapping variations in the strength and breadth of biogeographic transition zones using species turnover, *Proc. R. Soc. B*, *263*, 579–588.

Williams, P. H., H. M. de Klerk, and T. M. Crowe (1999), Interpreting biogeographical boundaries among Afrotropical birds: Spatial patterns in richness gradients and species replacement, *J. Biogeogr.*, *26*, 459–474.

Wilson, M. V., and A. Shmida (1984), Measuring beta diversity with presence–absence data, *J. Ecol.*, *72*, 1055–1064.

Yue, J. C., and M. K. Clayton (2005), A similarity measure based on species proportions, *Commun. Stat. -Theory Methods*, *34*, 2123–2131.