Quantifying biodiversity and asymptotics for a sequence of random strings

Hitoshi Koyano* and Hirohisa Kishino
Graduate School of Agricultural and Life Sciences, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan
(Received 27 April 2009; revised manuscript received 21 March 2010; published 7 June 2010)

We present a methodology for quantifying biodiversity at the sequence level by developing the probability theory on a set of strings. Further, we apply our methodology to the problem of quantifying the population diversity of microorganisms in several extreme environments and digestive organs and reveal the relation between microbial diversity and various environmental parameters.

DOI: 10.1103/PhysRevE.81.061912 PACS number(s): 87.23.−n, 02.50.Cw, 02.50.Tt

I. INTRODUCTION

The measurement of diversity or concentration within a population is a relevant issue across many areas and has been extensively studied since [1] (for example, see [2] for a classical bibliography on this subject). Biodiversity is one of the fundamental concepts in the study of ecology, and considerable research has been conducted on the methods for measuring this quantity since the early studies on the logarithmic-normal model [3,4]. The most widely used indices in the diversity analysis of animal and plant ecology are the Shannon index [5,6] and Simpson index [7] (for example, see [8,9] for a review on biodiversity indices). Theoretical and methodological studies on the Shannon index were conducted by [10–13] and others. The theory of the Simpson index was enriched by [14–19] and others. Historically, biodiversity has been measured using the species as the fundamental unit of analysis. Consequently, the problem of defining a biodiversity index is considered as the problem of constructing a function with the desired properties defined on the probability simplex $U=\{\boldsymbol{p}=(p_1,\ldots,p_n):p_i\geq 0,\ p_1+\cdots+p_n=1\}$ and taking values in the set of positive numbers when a set of entities divided into n categories is given. It has been pointed out that the above indices do not take into account the differences among categories [20–23].

Therefore, [21] defined a new diversity index called the quadratic entropy (see also [24–26]). Although the quadratic entropy is a diversity index developed under the framework stated above, with this index, a matrix of dissimilarities among categories can be chosen according to the objective of the research [27,28]. This index has been used for ecological analysis by [29,30] and others. Nucleotide diversity, which has been used in the study of population genetics and molecular evolution [31–34], is an index similar to quadratic entropy. This index has been applied in marine biology and microbial ecology by [35–38], by using taxonomic distance and sequence distance (the Hamming distance [39] in information theory) as dissimilarity measures, respectively. Quadratic entropy has several interesting properties and has led to the development of analysis of diversity as a generalization of analysis of variance. However, it is known that biodiversity is sometimes maximized when several species are eliminated as the difficulty of this index in conservation biology, which is one of the important fields in which biodiversity analysis is applied [40–42].

Moreover, because the Shannon and Simpson indices require clear definition of the species and unambiguous identification of each individual, it is difficult to apply these indices to microbial communities [43,44]. The quantification of microbial diversity is a basic requirement for biodiversity assessment because microorganisms occupy a large part of the total biomass on earth. However, the quantification of microbial diversity is difficult because of the above reasons, and several alternative methods have been explored [23,45,46]. See [47] for a recent review on the measurement of microbial diversity. Furthermore, [48] developed the phylogenetic diversity using a framework that was different from that of construction of a function on a probability simplex stated above. See [49–51] for theoretical studies on this index and [52–54] for the application of this index in conservation biology and microbial ecology.

In this study, we propose a methodology for quantifying biodiversity with sequence data, which have been rapidly accumulated in recent years. Because it is physically impossible to collect samples of all ribosomal RNA gene sequences in an environment, we have no option but to estimate the population diversity on the basis of one sample. Thus, we have two problems of the definition of the population diversity and the estimation from a sample. Traditionally, richness and evenness are the most important aspects that a diversity index should reflect. However, when diversity is measured by considering the divergence between all pairs of sequences, the evenness of categories becomes less important. Moreover, a sequence community is composed of several subcommunities, as described below, and therefore the diversity of a sequence community should be defined as the hierarchical quantity that reflects both the diversity within each subcommunity and the diversity between different subcommunities. For example, in calculating the average divergence between two randomly chosen sequences in a population, the aspect of hierarchy formation within a sequence community is not considered. Therefore, we define the population diversity of a sequence community as the quantity that reflects richness and hierarchy. We then propose a methodology for estimating population diversity and describe the theoretical foundation of this methodology.

In this paper, we deal with the problem of measuring biodiversity in the following framework. Letting A^* be the set of all strings on the set $\bar{A}$ composed of four letters a, g, c,

*h-koyano@zc4.so-net.ne.jp

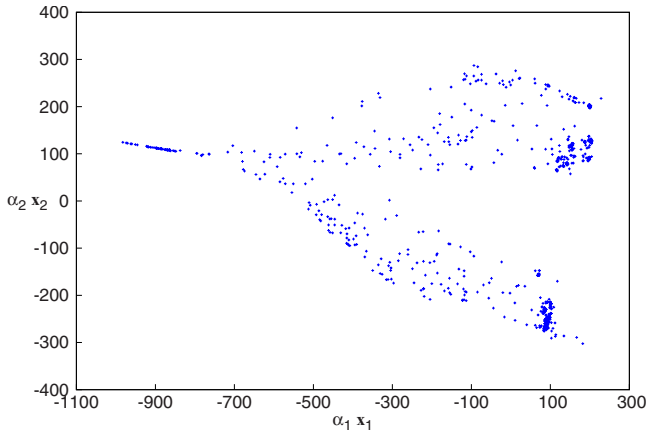

FIG. 1. (Color online) Graphical visualization of a sequence distribution 1. MDS for the environmental sample from the hot springs in Yellowstone National Park. α_1 and α_2 are the greatest and the second greatest eigenvalues of the matrix $A=(a_{ij})$, respectively, and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the eigenvectors for α_1 and α_2, respectively, where $a_{ij}=-[d_L(s_i,s_j)^2-\Sigma_{j=1}^{1068}d_L(s_i,s_j)^2/1068-\Sigma_{i=1}^{1068}d_L(s_i,s_j)^2/1068+\Sigma_{i=1}^{1068}\Sigma_{j=1}^{1068}d_L(s_i,s_j)^2/1068^2]/2$ for the sequences $s_1,\ldots,s_{1068}$ in the environmental sample (the sample size=1068).

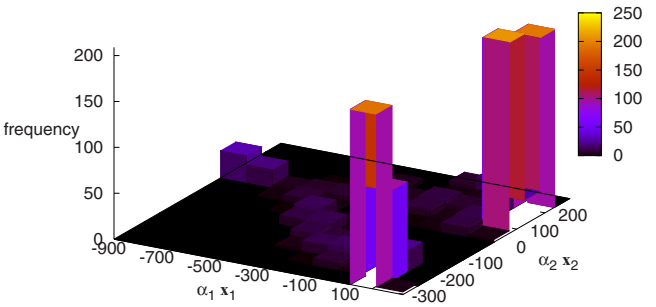

FIG. 2. (Color online) Graphical visualization of a sequence distribution 2. Histogram for the environmental sample from the hot springs in Yellowstone National Park. The labels of the first and the second coordinate axes $\alpha_1\boldsymbol{x}_1$ and $\alpha_2\boldsymbol{x}_2$ are the same as those in Fig. 1.

and t and the empty letter e, we construct a distance space (A^*,d_L) for the Levenshtein distance d_L [55]. The Levenshtein distance between two strings s and t is the minimal number of deletions, insertions, or substitutions required to transform s into t. Because A^* forms a noncommutative monoid by concatenation but is not a vector space, a mean for $s_1,\ldots,s_n\in A^*$ cannot be defined. Therefore, taking the consensus sequence [denoted by $\boldsymbol{m}(s_1,\ldots,s_n)$] as a measure of location, we can define a naive measure of dispersion as, for example,

$$\frac{1}{n}\sum_{i=1}^{n} d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n)) \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n))^2.$$

In this study, we approach the problem of quantifying biodiversity with ribosomal RNA gene sequences by developing a probability theory on A^*. We apply our methodology to quantify the diversity of all microorganisms in an environment, which has so far been difficult because of their enormous numbers. We then investigate the relationship between environmental parameters and diversity.

II. GRAPHICAL VISUALIZATION OF A SEQUENCE COMMUNITY

Figures 1 and 2 are the graphical visualizations of an environmental sample of microbial ribosomal RNA gene sequences collected from hot springs in Yellowstone National Park in the United States [56,57]. These figures were constructed in the following manner: (1) A distance matrix was computed for all the sequences in an environmental sample. The Levenshtein distance was chosen as the distance between two sequences. (2) Multidimensional scaling (MDS) [58] was applied to set the sequences in a plane (Fig. 1). (3) The plane was divided into several classes of rectangles, and the number of sequences belonging to each class was counted. A histogram of this information was then constructed. The number of classes was determined with Sturges' formula (Fig. 2). From these figures, we can observe that the distribution of the sequence community is not uniform but uneven, and not unimodal but multimodal. Figure 3 is the unrooted tree constructed from the same environmental sample. This figure indicates that the sequence community is classified into several subcommunities. On the basis of these observations, we model a sequence community in an environment as the mixture of several sequence groups having different unimodal distributions.

III. FRAMEWORK OF THE DIVERSITY ESTIMATION PROBLEM

Given n observations $x_1,\ldots,x_n\in\mathbb{R}$, their dispersion is often measured by the mean deviation or the variance

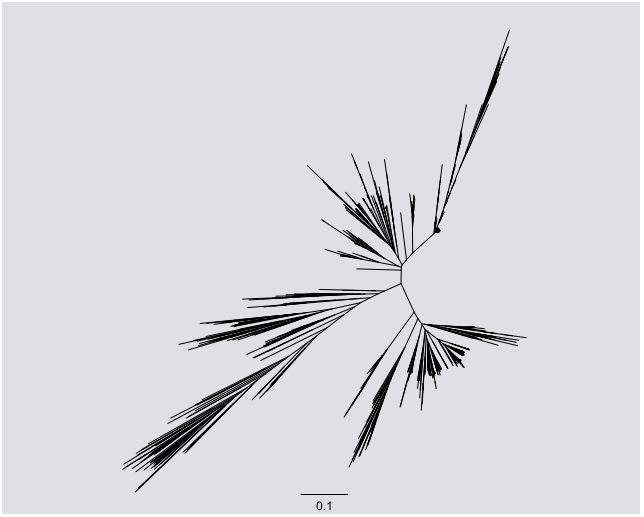

FIG. 3. (Color online) Graphical visualization of a sequence distribution 3. Unrooted tree for the environmental sample from the hot springs in Yellowstone National Park. The scale bar represents 0.1 substitution per site.

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|, \quad \frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|^2,$$

where $\bar{x}$ is the mean of $x_1,\ldots,x_n$. The basic idea for quantifying biodiversity with ribosomal RNA gene sequences in this study is to introduce an analogy of the above measures of dispersion for sequences $s_1,\ldots,s_n \in A^*$ by substituting the consensus sequence and the Levenshtein distance for the mean and the Euclidean distance, respectively. However, because an environmental sample of ribosomal RNA gene sequences is multimodal as shown in Fig. 2, the naive counterpart of the mean deviation or the variance defined in the manner stated above is defective as a measure of dispersion. Consequently, the layering of variances which are described below is required. Further, because it is impossible to collect all ribosomal RNA gene sequences present in a particular environment, an environmental sample of ribosomal RNA gene sequences can be regarded as a random sample drawn from the population of ribosomal RNA gene sequences in this environment. To represent this probabilistic structure of the problem of quantifying biodiversity, the concept of a random string is introduced in the following paragraphs and in the supplementary material [59].

In this section, we state our framework of the problem of estimating biodiversity and the theoretical results on the accuracy of our estimator. Precise definitions and rigorous proofs of the concepts and propositions stated here are presented in Sec. II of the supplementary material [59]. Roughly, the setting of the estimation problem is as follows: (i) We observe n sequences $s_1,\ldots,s_n$ in an environment. (ii) s_i belongs to one of k groups for each $i=1,\ldots,n$, and all the sequences in the jth group are generated from an identical distribution for each $j=1,\ldots,k$. (iii) The distribution of each group as well as the number k of groups and the group to which each s_i belongs are unknown. (iv) In this situation, we estimate the diversity of all the sequences in the environment.

First, we introduce the following statistics. In this study, the mean of the Levenshtein distances from $s_1,\ldots,s_n \in A^*$ to their consensus sequence,

$$\boldsymbol{v}(s_1,\ldots,s_n)=\frac{1}{n}\sum_{i=1}^{n} d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n)), \tag{1}$$

is termed the variance of $s_1,\ldots,s_n$, although it is an analogy of the mean deviation. The normalized variance is defined as

$$\boldsymbol{w}(s_1,\ldots,s_n)=\frac{1}{n}\sum_{i=1}^{n}\frac{d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n))}{\max\{|s_i|,|\boldsymbol{m}(s_1,\ldots,s_n)|\}}.$$

We have $0\le \boldsymbol{w}(s_1,\ldots,s_n)\le 1$ from $d_L(s,t)\le \max\{|s|,|t|\}$. Because the sequences in an environmental sample are divided into a number of groups, as shown in Fig. 3, the consensus sequence of all the sequences in a sample is not an appropriate measure of the center of the sample. Consequently, we cannot apply these statistics to estimate the diversity of the population of a sequence community in an environment.

We, therefore, consider the following statistics, which are generalizations of the above. Let us suppose that a sample is divided into k groups $C_1,\ldots,C_k$. Let n_i stand for the size of the ith group C_i for each $i=1,\ldots,k$ and s_{ij} be the jth sequence in C_i for each $j=1,\ldots,n_i$. Then, $C_i=\{s_{i1},\ldots,s_{in_i}\}$. The consensus sequence and variance of the sequences in C_i are denoted by $\boldsymbol{m}(C_i)$ and $\boldsymbol{v}(C_i)$, respectively. The distance via $\boldsymbol{m}(C_i)$ between the consensus sequence $\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k))$ of the consensus sequences of k groups and the jth sequence s_{ij} in the ith group is given by $d_L(\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k)),\boldsymbol{m}(C_i))+d_L(\boldsymbol{m}(C_i),s_{ij})$. Averaging the above quantities with respect to i and j, we get

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}\{d_L(\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k)),\boldsymbol{m}(C_i))+d_L(\boldsymbol{m}(C_i),s_{ij})\}$$

$$=\boldsymbol{v}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k))+\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{v}(C_i). \tag{2}$$

This is denoted by $\boldsymbol{v}_2(s_1,\ldots,s_n)$ and is termed the two-layer variance. The variance (1) is equal to Eq. (2) with $k=1$, where the sample consists of only one group. Then,

$$\boldsymbol{w}_2(s_1,\ldots,s_n)=\boldsymbol{w}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k))+\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{w}(C_i) \tag{3}$$

is termed the normalized two-layer variance. We have $0\le \boldsymbol{w}_2(s_1,\ldots,s_n)\le 2$. Although we can define an l-layer variance for any $l\ge 1$, repeating the division of each group into several subgroups, in this study, we use the two-layer variance. A method for grouping sequences will be considered later.

We next formulate the problem of statistically estimating the diversity of a sequence community and state an asymptotic property of the estimator given in Eq. (2).

(1) Random letters. We construct a measurable space $(\bar{A},2^{\bar{A}})$ for the power set $2^{\bar{A}}$ of $\bar{A}$ and define a random letter as a random variable defined on a probability space $(\Omega,\mathfrak{F},P)$ and taking values in $(\bar{A},2^{\bar{A}})$. The distribution and probability function of a random letter α are defined as

$$Q(B)=P(\{\omega\in\Omega:\alpha(\omega)\in B\}), \quad B\in 2^{\bar{A}},$$

$$q(x)=Q(\{x\}), \quad x\in\bar{A},$$

respectively. A letter $x^*\in\bar{A}$ satisfying $q(x)\le q(x^*)$ for any $x\in\bar{A}$ is called a consensus letter of α and is denoted by $m'(\alpha)$. Hereafter, we consider only a random letter for which there exists a unique consensus letter.

(2) Random strings. Let us denote a set of finite sequences of the elements of A to which the infinite sequence $(e,\ldots)$ of the empty letter is appended by A^*. We construct a measurable space $(A^*,2^{A^*})$ for the power set 2^{A^*} of A^* and define a random string as a sequence $\sigma=\{\alpha_n:n\in\mathbb{Z}^+\}$ ($\mathbb{Z}^+$ is the set of all positive integers) of random letters, which satisfies the following conditions:

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

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

Because, in general, the length of a string $\sigma(\omega) \in A^*$ depends on $\omega \in \Omega$, we treat a random string as a stochastic process with a discrete time parameter and not a random vector with each element A valued. A consensus sequence of a random string $\sigma=\{\alpha_n : n \in \mathbb{Z}^+\}$ is defined as $\boldsymbol{m}'(\sigma) = \{m'(\alpha_n) : n \in \mathbb{Z}^+\}$. We set

$$\boldsymbol{v}'(\sigma) = \sum_{s \in A^*} d_L(s, \boldsymbol{m}'(\sigma)) \boldsymbol{q}_{1,\ldots,|s|+1}(x_1, \ldots, x_{|s|}, e),$$

$$s = (x_1, \ldots, x_{|s|}, e, \ldots),$$

for a probability function $\boldsymbol{q}_{1,\ldots,|s|+1}$ of a finite-dimensional distribution of $\sigma=\{\alpha_n : n \in \mathbb{Z}^+\}$ at sites $1, \ldots, |s|+1$ for each $s \in A^*$ and call $\boldsymbol{v}'(\sigma)$ a variance of σ.

(3) Estimation problem. Let us suppose that we observe n sequences in an environment, and they are divided into k groups,

$$C_1 = \{s_{11}, \ldots, s_{1n_1}\}, \ldots, C_k = \{s_{k1}, \ldots, s_{kn_k}\}, \quad \sum_{i=1}^{k} n_i = n.$$

We suppose that s_{ij} $(1 \le i \le k, \ 1 \le j \le n_i)$ is a realization of a random string defined on a probability space $(\Omega, \mathfrak{F}, P)$ and taking values in $(A^*, 2^{A^*})$ and $s_{i1}, \ldots, s_{in_i}$ $(1 \le i \le k)$ are realizations of random strings $\sigma_{i1}, \ldots, \sigma_{in_i}$, which are independent and have identical finite-dimensional distributions. Let $\boldsymbol{m}_i'$ and $\boldsymbol{v}_i'$ be the consensus sequence and variance of σ_{i1}, respectively, for each $i=1, \ldots, k$. We introduce a quantity

$$\boldsymbol{v}(\boldsymbol{m}_1', \ldots, \boldsymbol{m}_k') + \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{v}_i', \tag{4}$$

combining the dispersion between communities and the variance within each community and formulate the problem of estimating biodiversity as the problem of estimating Eq. (4). The diversity (4) is a known function of unknown parameters $\boldsymbol{m}_i'$ and $\boldsymbol{v}_i'$ $(1 \le i \le k)$. We estimate these parameters by $\boldsymbol{m}(C_i)$ and $\boldsymbol{v}(C_i)$ $(1 \le i \le k)$, respectively, which is equivalent to estimating Eq. (4) by the two-layer variance (2). Under certain regular conditions, $\boldsymbol{m}(C_i)$ and $\boldsymbol{v}(C_i)$, which depend on n_i, almost surely converge to $\boldsymbol{m}_i'$ and $\boldsymbol{v}_i'$, respectively, as $n_i \to \infty$ for each $i=1, \ldots, k$. Hence, Eq. (4) is consistently estimated by Eq. (2). These asymptotic results are demonstrated in Theorems 2 and 3 in Sec. II of the supplementary material [59].

IV. SEQUENCE CLASSIFICATION

Next, we consider the problem of classifying sequences. We were unable to use the conventional nonhierarchical cluster analysis or discriminant analysis because it was difficult to determine the number of sequence subcommunities from Fig. 1 or Fig. 3. Therefore, we classified the sequences according to the following algorithm. In this algorithm, groups of sequences are formed by collecting sequences for which neighborhoods intersect.

Let S be a list of sequences. We define lists $U_\varepsilon(s)$ and $\mathcal{U}_\varepsilon$ as

$$U_\varepsilon(s) = \langle t \in S | d_L(t,s) < \varepsilon \rangle, \quad \mathcal{U}_\varepsilon = \langle U_\varepsilon(s) | s \in S \rangle$$

for $\varepsilon > 0$ and $s \in S$, respectively. $\mathcal{U}_\varepsilon$ is a nested list. We denote the number of elements and the ith element of a list L by length (L) and $L[i]$, respectively, and use the symbols of set operations for the lists,

```
for i ← 1 to length (U_ε) − 1
  do for j ← i + 1 to length (U_ε)
    do if U_ε[i] ∩ U_ε[j] ≠ ∅
      then U_ε[i] ← U_ε[i] ∪ U_ε[j]
           U_ε ← U_ε − U_ε[j]
           j ← i
```

By this procedure, $\mathcal{U}_\varepsilon$ becomes the direct sum decomposition of S. Although we can apply the algorithm corresponding to the k means [60] by replacing the mean vectors with consensus sequences, we mention the following as the characteristics of our algorithm, especially in contrast to the k means:

(1) For executing the algorithm, we must set the radius ε of the neighborhood instead of the number of clusters. We set the value of ε by considering the following two points: (i) sequences divided into a subcommunity are classified as an identical category from the perspective of genomics, such as homology search and (ii) after dividing, the sequences in a subcommunity are distributed such that they center around the consensus sequence.

(2) Our algorithm has the following two advantages: (i) it is independent of an initial division and an order of sequences and (ii) the distance from any sequence to the nearest in a subcommunity cannot be remarkably great.

There is no theoretical assurance that this algorithm classifies sequences correctly in some framework and, therefore, we check the result of executing the algorithm in the following two manners:

(1) After classification, we construct the consensus sequences for sequence subcommunities and compare the result of the MDS applied to their set with that applied to the original environmental sample of sequences (Fig. 4).

(2) We constructed a tree by combining the original environmental sample and the set of consensus sequences, and we check the distribution of the consensus sequences on the tree (Fig. 5).

By comparing Figs. 1 and 4, we find that the set of sequences is made equally sparse throughout the algorithm. Further, the consensus sequences are distributed almost uniformly on the tree in Fig. 5. Although the above inspection is not complete because the correspondence between the Levenshtein distance on A^* and the distance on $\mathbb{R}^2$ is not perfect, and the Levenshtein distance between two sequences was transformed into the Levenshtein distance per site in the construction of the tree, from these observations, it seems that our algorithm basically classified the sequences as expected.

V. ROBUSTNESS WITH RESPECT TO THE SAMPLE SIZE

With respect to the accuracy of our diversity estimator, besides the regular conditions on mathematical statistics and the classification of sequences, we encounter an instability problem because considering the enormous numbers of microorganisms in an environment, an environmental sample

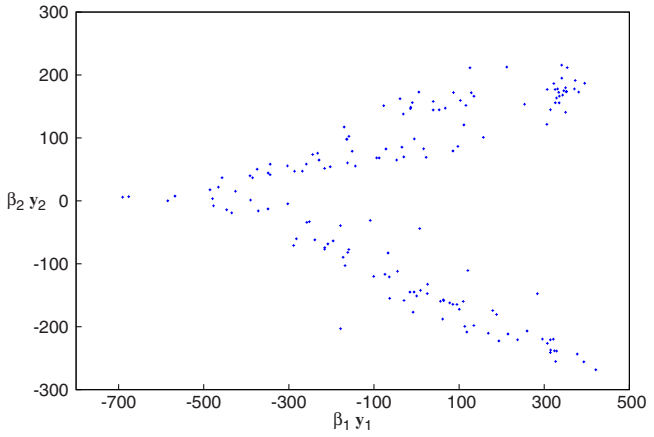

FIG. 4. (Color online) Checking the performance of the sequence classification algorithm 1. MDS for the consensus sequences. β_1 and β_2 are the greatest and the second greatest eigenvalues of the matrix $B=(b_{ij})$, respectively, and $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the eigenvectors for β_1 and β_2, respectively, where $b_{ij}=-\{d_L(\boldsymbol{m}(C_i),\boldsymbol{m}(C_j))^2-\Sigma_{j=1}^{158}d_L(\boldsymbol{m}(C_i),\boldsymbol{m}(C_j))^2/158-\Sigma_{i=1}^{158}d_L(\boldsymbol{m}(C_i),\boldsymbol{m}(C_j))^2/158+\Sigma_{i=1}^{158}\Sigma_{j=1}^{158}d_L(\boldsymbol{m}(C_i),\boldsymbol{m}(C_j))^2/158^2\}/2$ for the consensus sequences $\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_{158})$ for 158 sequence subcommunities (the number of subcommunities=158).

may not be collected from all over the subcommunities of 16S ribosomal RNA gene sequences in the environment. Therefore, we randomly draw subsamples sized 200, 400, 600, and 800 from the environmental sample sized 1068 obtained from the hot springs. We then plot the total branch length and average branch length of the trees, and the two-layer variance and the normalized two-layer variance for these subsamples in Figs. 6 and 7, respectively. From these figures, we can see that while the length of the tree increases and the average branch length shortens with the increase in the sample size, the hierarchical variances have a stable transition.

VI. APPLICATION TO MICROBIAL DIVERSITY ESTIMATION

We estimate the diversity of the populations of microorganisms by applying the methodology described above to the environmental samples of microbial 16S ribosomal RNA gene sequences collected from the following nine environments: hot springs, alkaline lakes, hypersaline lagoons, Antarctica, deep-sea trenches, deep-sea vent fields, the human mouth, the human stomach, and the cow's rumen. For comparison, we also apply our methodology to the environmental samples of microbial whole-genome shotgun sequences from the following two environments: the Sargasso Sea and the human intestine. The detailed information on these data is given in Sec. I of the supplementary material [59].

The length of sequences varies depending on the environment as shown in Table I, and it is unreasonable to assume that its variation is derived from microbiota in each environment. Although an environmental sample is not a sample obtained by collecting only the target sequences determined before sampling, it is reasonable to assume that the environmental samples we analyzed were independently drawn not from the set of all 16S ribosomal RNA gene sequences in the environment but from the set of sequences with the length in a certain range depending on the environment. This shows that the independence of sequences that is one of the regular conditions for the strong consistency of the diversity estimator is not completely satisfied (see Theorems 2 and 3 in Sec. II of the supplementary material [59]). Consequently, we use the normalized two-layer variance $\boldsymbol{w}_2(s_1,\ldots,s_n)$ (hereafter N2LV) for comparing the diversity between the environments. This is an estimate for the normalized version,

$$\boldsymbol{w}(\boldsymbol{m}'_1,\ldots,\boldsymbol{m}'_k)+\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{w}'_i,$$

of the diversity (4) in the interval [0,2], where $\boldsymbol{w}'_i$ is the quantity defined by replacing the Levenshtein distance in $\boldsymbol{v}'_i$ with the Levenshtein distance per site and is therefore free from the attribute of length.

The estimation results are shown in Table I. First, we examine the results of the application of our methodology to

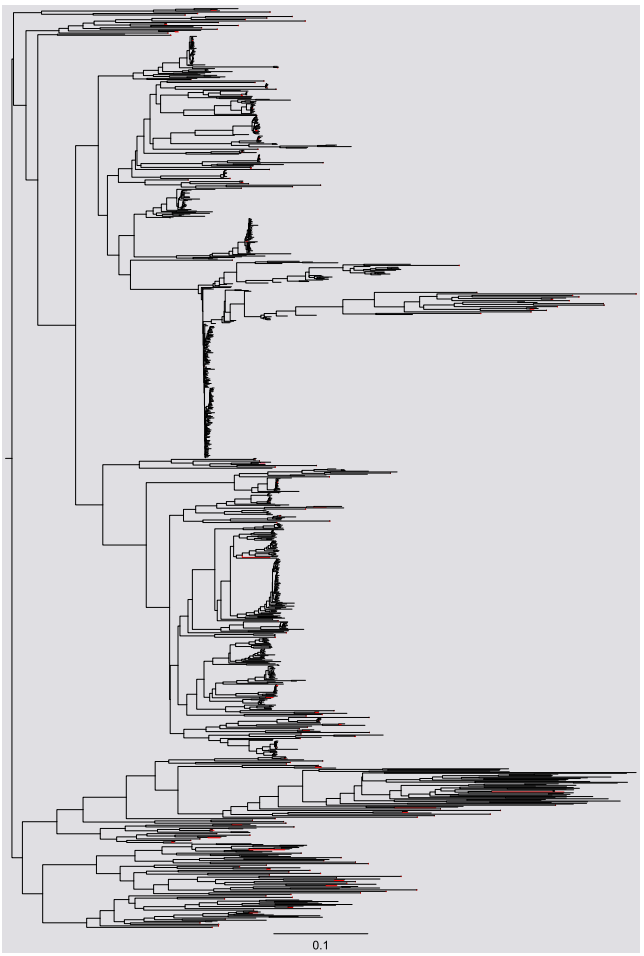

FIG. 5. (Color online) Checking the performance of the sequence classification algorithm 2. The distribution of the consensus sequences on the tree. The red (light gray) branches of the tree stand for the consensus sequences. The scale bar represents 0.1 substitution per site.

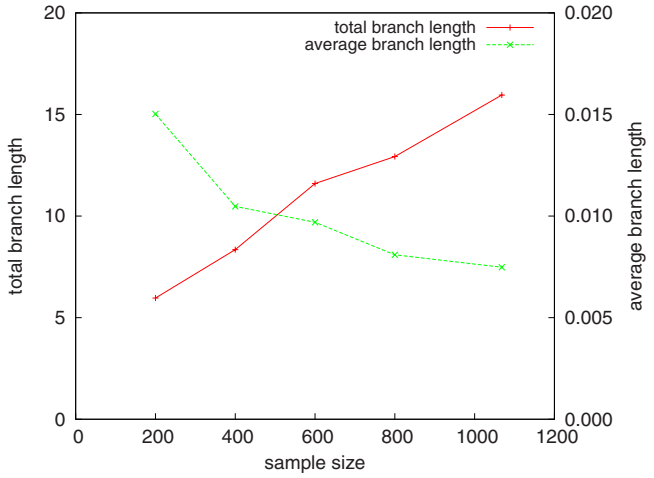

FIG. 6. (Color online) Inspecting the robustness of the diversity estimator to the sample size 1. Plotting of the total and average branch lengths of the trees on the basis of the subsamples sized 200, 400, 600, and 800.

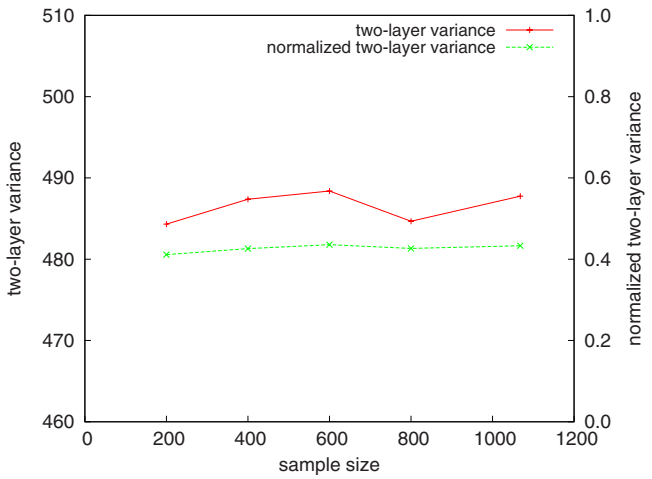

FIG. 7. (Color online) Inspecting the robustness of the diversity estimator to the sample size 2. Plotting of the two-layer variance and the normalized two-layer variance on the basis of the subsamples sized 200, 400, 600, and 800.

the environmental samples of whole-genome shotgun sequences from the Sargasso Sea and the human intestine. An environmental sample of whole-genome shotgun sequences is composed of miscellaneous and incommensurable sequences and, therefore, the computation results of the two-layer variance obtained by applying our methodology to this sample cannot necessarily be the quantity that reflects biodiversity, unlike for an environmental sample of 16S ribosomal RNA gene sequences. However, the point here is that judging from the characteristics of the two kinds of samples, the maximum of the two-layer variance for an environmental sample of 16S ribosomal RNA gene sequences cannot exceed that for an environmental sample of whole-genome shotgun sequences. The N2LV of both environments is approximately 0.56, and we can infer that the saturation point of the N2LV is around this level when an enormous number of sequences are randomly collected in an environment, although the N2LV lies in the interval [0,2]. The N2LV values do not depend on the sample size. After trying 5%, 10%, and 20% of the average length of sequences in a sample as the neighborhood radius for grouping sequences, we adopted the value of 10%.

TABLE I. Estimates of microbial diversity. Hot spring (HS), alkaline lake (AL), hypersaline lagoon (HL), Antarctica (AT), deep-sea trench (DT), deep-sea vent field (DV), Sargasso Sea (SS), human mouth (M), human stomach (S), human intestine (I), and cow's rumen (CR).

Environment	HS	AL	HL	AT	DT	DV
Sample size	1068	558	1655	1056	441	800
Max. length of sequence	1471	1551	1512	1563	1646	2031
Min. length of sequence	188	247	296	267	385	241
Average length of sequence	1212.79	769.47	1020.43	875.82	1199.52	1028.43
Number of subcommunities	158	286	580	183	214	333
Max. size of subcommunity	234	61	113	452	37	147
Two-layer variance	487.75	372.06	419.93	403.73	477.19	495.25
Normalized two-layer variance	0.4332	0.4180	0.3705	0.2843	0.3490	0.4156

Environment	SS	M	S	I	CR
Sample size	1500	1186	111	1500	667
Max. length of sequence	4653	1627	1511	4976	1550
Min. length of sequence	103	206	434	97	577
Average length of sequence	1101.84	777.65	724.64	1436.37	883.29
Number of subcommunities	1447	258	34	989	144
Max. size of subcommunity	51	463	33	153	139
Two-layer variance	674.03	498.02	374.16	904.74	505.37
Normalized two-layer variance	0.5629	0.3525	0.3919	0.5661	0.3967

First, we examine the results for the extreme environments. Although innumerable parameters are required for describing an environment, we consider five parameters, namely, (i) temperature, (ii) hydrogen ion concentration, (iii) salt concentration, (iv) pressure, and (v) partial pressure of oxygen. We then study how the N2LV change with variations in these five environmental parameters to identify the parameter that has the greatest influence on the diversity. The low-temperature environments of Antarctica and the deep-sea trenches had a smaller N2LV than the other environments. Conversely, the N2LVs of the high-temperature environments of the hot springs and the deep-sea vent fields were fairly large. The N2LVs of the alkaline lakes and the hypersaline lagoon were also comparatively large. These observations suggest that among the five environmental parameters, temperature has the greatest influence on microbial diversity and, specifically, low temperature markedly reduces the variety of microbes.

With regard to the digestive organs, the intestine had a large N2LV, which suggests the existence of very diverse microbes. The N2LV for the mouth sample was relatively small, from which we inferred that there were relatively less diverse microbes in the mouth, despite the presence of a large number of microorganisms. The N2LVs of the human stomach and the cow's rumen were very close, suggesting that these two environments have a similar level of microbial diversity, although they have very different microbiota.

Many metagenomic researches have been performed in recent times, and large-scale environmental samples of whole-genome shotgun sequences have been accumulated (for example, see [61–63]). Although these samples are a huge set of sequences collected quite randomly and no methodology with a rigorous theoretical basis is available for dealing with them, we wish to extend our methodology in order to perform microbial diversity analysis using these samples.

[1] C. Gini, Studi Economico-Giuridici dell Universiti di Cagliari **3**, 1 (1912).
[2] B. Dennis, G. P. Patil, O. Rossi, S. Stehman, and C. Taillie, in *Ecological Diversity in Theory and Practice*, edited by J. F. Grassle, G. P. Patil, W. K. Smith, and C. Taillie (International Cooperative Publishing House, Fairland, MD, 1979), Vol. 1, pp. 319–354.
[3] R. A. Fisher, A. S. Corbet, and C. B. Williams, J. Anim. Ecol. **12**, 42 (1943).
[4] F. W. Preston, Ecology **29**, 254 (1948).
[5] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[6] C. E. Shannon, Bell Syst. Tech. J. **27**, 623 (1948).
[7] E. H. Simpson, Nature (London) **163**, 688 (1949).
[8] J. A. Ludwig and J. F. Reynolds, *Statistical Ecology: A Primer on Methods and Computing* (Wiley-Interscience, New York, 1988).
[9] A. E. Magurran, *Measuring Biological Diversity* (Blackwell, Oxford, 2004).
[10] K. Hutcheson, J. Theor. Biol. **29**, 151 (1970).
[11] K. O. Bowman, K. Hutcheson, E. P. Odum, and L. R. Shenton, in *International Symposium on Statistical Ecology*, edited by G. P. Patil, E. C. Pielou, and W. E. Waters (Pennsylvania State University Press, University Park, PA, 1971), Vol. 3, pp. 315–359.
[12] N. I. Lyons, Am. Nat. **118**, 438 (1981).
[13] W. B. Sherwin, F. Jabot, R. Rush, and M. Rossetto, Mol. Ecol. **15**, 2857 (2006).
[14] W. J. Ewens, Theor Popul. Biol. **3**, 87 (1972).
[15] M. Kimura and T. Ohta, Proc. Natl. Acad. Sci. U.S.A. **72**, 2761 (1975).
[16] M. Nei, T. Maruyama, and R. Chakraborty, Evolution (Lawrence, Kans.) **29**, 1 (1975).
[17] B. S. Weir and C. C. Cockerham, Evolution (Lawrence, Kans.) **38**, 1358 (1984).
[18] D. E. Pearse and K. A. Crandall, Conserv. Genet. **5**, 585 (2004).
[19] R. S. Etienne and H. Olff, Ecol. Lett. **7**, 170 (2004).
[20] M. Nei and W. H. Li, Proc. Natl. Acad. Sci. U.S.A. **76**, 5269 (1979).
[21] C. R. Rao, Theor Popul. Biol. **21**, 24 (1982).
[22] S. H. Cousins, Trends Ecol. Evol. **6**, 190 (1991).
[23] M. G. Watve and R. M. Gangal, Appl. Environ. Microbiol. **62**, 4299 (1996).
[24] C. R. Rao, Sankhya, Ser. A **44**, 1 (1982).
[25] C. Rao and T. Nayak, IEEE Trans. Inf. Theory **31**, 589 (1985).
[26] C. R. Rao, in *Encyclopedia of Statistical Sciences*, edited by S. Kotz and N. L. Johnson (Wiley, New York, 1986), Vol. 7, pp. 614–617.
[27] A. W. F. Edwards, Biometrics **27**, 873 (1971).
[28] M. Nei, Am. Nat. **106**, 283 (1972).
[29] J. Izsák and L. Papp, Environ. Ecol. Stat. **2**, 213 (1995).
[30] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman, Science **308**, 1635 (2005).
[31] M. Kimura, Genet. Res. **11**, 247 (1968).
[32] M. Nei and F. Tajima, Genetics **97**, 145 (1981).
[33] F. Tajima, Genetics **105**, 437 (1983).
[34] M. Nei and L. Jin, Mol. Biol. Evol. **6**, 290 (1989).
[35] R. M. Warwick and K. R. Clarke, Mar. Ecol.: Prog. Ser. **129**, 301 (1995).
[36] R. M. Warwick and K. R. Clarke, J. Appl. Ecol. **35**, 532 (1998).
[37] K. R. Clarke and R. M. Warwick, J. Appl. Ecol. **35**, 523 (1998).
[38] A. P. Martin, Appl. Environ. Microbiol. **68**, 3673 (2002).
[39] R. W. Hamming, Bell Syst. Tech. J. **29**, 147 (1950).
[40] S. Pavoine, S. Ollier, and D. Pontier, Theor Popul. Biol. **67**, 231 (2005).
[41] K. Shimatani, Oikos **93**, 135 (2001).
[42] J. Izsák and L. Szeidl, Environ. Ecol. Stat. **9**, 423 (2002).
[43] J. T. Staley, in *Microbiology-1980*, edited by D. Schlessinger (American Society for Microbiology, Washington, D.C., 1980), pp. 321–322.
[44] V. Torsvik, K. Salte, R. Sorheim, and J. Goksoyr, Appl. Envi-

ron. Microbiol. **56**, 776 (1990).
[45] S. H. Hong, J. Bunge, S. O. Jeon, and S. S. Epstein, Proc. Natl. Acad. Sci. U.S.A. **103**, 117 (2006).
[46] J. B. Hughes, J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan, Appl. Environ. Microbiol. **67**, 4399 (2001).
[47] C. A. Lozupone and R. Knight, FEMS Microbiol. Rev. **32**, 557 (2008).
[48] D. P. Faith, Biol. Conserv. **61**, 1 (1992).
[49] M. Steel, Syst. Biol. **54**, 527 (2005).
[50] F. Pardi and N. Goldman, PLoS Genet. **1**, e71 (2005).
[51] K. Hartmann and M. Steel, Syst. Biol. **55**, 644 (2006).
[52] P. Posadas, D. R. M. Esquivel, and J. V. Crisci, Conserv. Biol. **15**, 1325 (2001).
[53] F. Forest, R. Grenyer, M. Rouget, T. J. Davies, R. M. Cowling, D. P. Faith, A. Balmford, J. C. Manning, S. Proches, M. van der Bank, G. Reeves, T. A. J. Hedderson, and V. Savolainen, Nature (London) **445**, 757 (2007).
[54] C. A. Lozupone and R. Knight, Proc. Natl. Acad. Sci. U.S.A. **104**, 11436 (2007).
[55] V. I. Levenshtein, Sov. Phys. Dokl. **10**, 707 (1966).
[56] S. E. Korf, W. P. Inskeep, R. E. Macur, M. A. Kozubal, W. P. Taylor, and A. Nagy (unpublished).
[57] S. E. Korf, A. M. Nagy, R. E. Macur, M. A. Kozubal, W. P. Taylor, W. P. Inskeep, G. Ackerman, and D. Masur (unpublished).
[58] W. S. Torgerson, Psychometrika **17**, 401 (1952).
[59] See supplementary material at http://link.aps.org/supplemental/10.1103/PhysRevE.81.061912 for information on the material analyzed in this study and a detailed description of the statistical theory for biodiversity estimation.
[60] J. B. MacQueen, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. LeCam and J. Neyman (University of California Press, Berkeley, CA, 1967), Vol. 1, pp. 281–297.
[61] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson, Science **312**, 1355 (2006).
[62] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith, Science **304**, 66 (2004).
[63] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork, Science **315**, 1126 (2007).