A dynamic programming algorithm for binning microbial community profiles

Quansong Ruan1, Joshua A. Steele2, Michael S. Schwalbach2, Jed A. Fuhrman2 and Fengzhu Sun3,*

1Department of Mathematics, University of Southern California, 3620 South Vermont Avenue, KAP 108, Los Angeles, California 90089-253, USA, 2Department of Biological Sciences, University of Southern California, 3616 Trousdale Parkway, AHF 107, Los Angeles, CA 90089-0371, USA and 3Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, MCB 201, Los Angeles, CA 90089-2910, USA

ABSTRACT
Motivation: A number of community profiling approaches have been widely used to study the microbial community composition and its variations in environmental ecology. Automated Ribosomal Intergenic Spacer Analysis (ARISA) is one such technique. ARISA has been used to study microbial communities using 16S–23S rRNA intergenic spacer length heterogeneity at different times and places. Owing to errors in sampling, random mutations in PCR amplification, and probably mostly variations in readings from the equipment used to analyze fragment sizes, the data read directly from the fragment analyzer should not be used for downstream statistical analysis. No optimal data preprocessing methods are available. A commonly used approach is to bin the reading lengths of the 16S–23S intergenic spacer. We have developed a dynamic programming algorithm based binning method for ARISA data analysis which minimizes the overall differences between replicates from the same sampling location and time.

Results: In a test example from an ocean time series sampling program, data preprocessing identified several outliers which upon re-examination were found to be because of systematic errors. Clustering analysis of the ARISA from different times based on the dynamic programming algorithm binned data revealed important features of the biodiversity of the microbial communities.

Availability: The algorithm is implemented in a software package and it is available upon request from the corresponding author.

Contact: fsun@usc.edu

INTRODUCTION

A number of community profiling approaches have been widely used to study microbial community composition and its variation in environmental microbiology, both aquatic and terrestrial in the last decade. These methods include terminal restriction fragment polymorphism (TRFLP) (Avaniss-Aghajani et al., 1994; Liu et al., 1997), denaturing gradient gel electrophoresis (DGGE) (Muyzer et al., 1993; Troussellier et al., 2002) and Automated Ribosomal Intergenic Spacer Analysis (ARISA) (Fisher and Triplett, 1999). They are based on the fact that certain ribosomal RNA regions are highly heterogenous and well-conserved for different species during evolution.

ARISA is based on length heterogeneity of the intergenic regions between the 16S rRNA and 23S rRNA genes. The ARISA fragments (intergenic region + gene fragments) are mostly found between 400 and 1200 bp. After the DNA sample is extracted from the community, it is amplified by PCR with a fluorescent primer. Then the amplified sample (a mixture of fragments of different lengths) is analyzed by gel electrophoresis. Fragments of different sizes have different mobility in the gel. A laser detector records their mobility, and the amount of fluorescence indicates the relative abundance of fragments of that size. Thus, an electropherogram with peaks at different sizes is obtained for each community. Each peak is called an operational taxonomic unit (OTU), identified by its characteristic fragment size.

There are several sources of error in ARISA profiles. First, ARISA profiles show the strongest and most reproducible signal for the most abundant bacteria within a sample, with decreased reproducibility in the detection of bacteria present in low abundance near the detection limit. Second, the community sample is a random sample from the whole community of interest and different profiles may be obtained for different samples, depending on the scale of patchiness in the environment. Third, the PCR amplification process is stochastic and susceptible to error, hence errors in the length of the ITS may be introduced at a low rate. Fourth, the ABI machine used for fragment analysis determines fragment lengths with good precision but not always with good accuracy, with the inaccuracies increasing as fragment size increases (Brown et al., 2005). Therefore, two fragments of identical length could erroneously be sized differently or two fragments of different lengths could be sized the same across gels owing to the above mentioned accuracy issues. To address this issue a rigorous binning strategy is necessary.

Several empirical approaches have been suggested for binning the ARISA profiles. Hewson and Fuhrman, (2004) used a bin size of 3 for fragment length ≤500 bp and bin size of 7 for fragment length >500. Fisher and Triplett (1999) studied the size variations for different sizing standards. They found that the difference is about 1–2 bp for fragment size below 1 kb, 3–5 bp for fragment size of up to 1150 bp and the largest difference of 13 bp for fragment size of 1193 bp in one standard and 1180 bp in another standard. Based on experimental results, Brown et al. (2005) suggested that windows of 3 bp wide be used from 400 to 700 bp, windows of 5 bp...
wide from 701 to 1000 bp, and windows of 10 bp wide from 1000 to 1200 bp.

In addition to fixed size binning, i.e. the same bin size within certain range and the same binning method for all profiles or communities of interest, binning strategies based on comparison of replicates of each community have been developed (Hewson and Fuhrman, 2004; Moeseneder et al., 2001; Stepanauskas et al., 2003). Hewson and Fuhrman (2005) discussed these methods and further proposed another binning method in order to compare communities—they considered sliding window frames (of shift by 1 at a time) and looked at the mean and maximum pairwise similarity.

However, binning with fixed length within a certain range may not be optimal. Suppose that in a certain range, say 400–600 bp, the reading error can be +1 or −1. Intuitively we should bin the profiles using bin size 3. Suppose we have two OTUs with lengths, 419 and 423 bp, respectively, and no others in that range. It is obvious that the optimal binning method should be (418, 419, 420), (421), (422, 423, 424). However, no such binning method exists if we insist on binning of 3 bp at each bin. Therefore we should allow bins of variable sizes with sizes possibly less than some MaxBinSize (the MaxBinSize depends on the starting position of the bin).

Our goal is to construct a binning strategy with variable bin sizes so that same OTUs will be treated the same in different samples while different OTUs will be treated as different.

**APPROACHES**

In this paper, we develop a dynamic programming algorithm (DPA) for microbial community profile binning. Our approach differs from previous approaches in two main aspects. First, instead of binning using fixed sizes, the bin sizes vary and are determined from the data. Second, a DPA is used for binning to minimize the total difference between the replicates. The use of DPA makes it possible to find an optimal solution in a reasonable time (about 1 min on a 2.8 GHz PC) for our data (42 pairs of ARISA replicates of lengths ranging from 398 to 1200 bp) amid exponential number of possible solutions.

Although it has been found that the peak heights may not be proportional to the relative abundance of bacterial species (Suzuki and Giovannoni, 1996; Yannarell and Triplett, 2004), the profile replicates from the same sample spot are expected to be highly similar. The data from our laboratory show that the profiles from replicate samples are much more similar than the profiles from different samples, indicating the reproducibility of the data. To overcome the problem that the observed relative fluorescence of different peaks may be different from the relative abundance of organisms in the sample, we also use the presence/absence data derived from the observed profiles. An OTU is defined as present if the observed intensity is above a given threshold.

This paper is organized as follows. In the Materials and Methods section, we first describe our approach for data preprocessing. The basic idea for the score function of different binning approaches is then given. A DPA for profile binning is described in detail. The binning results using both the relative intensity data as well as presence/absence data are given in the Results section. We then briefly describe our efforts to use the binned data for downstream analysis and the biological implications. The paper concludes by describing some limitations of our approach and future directions for further research.

**MATERIALS AND METHODS**

We started with 47 pairs of replicate samples that had previously been analyzed using ARISA. They were sampled from different times ranging from August 2000 to December 2004 (roughly once per month) from the same geographical location—the San Pedro Channel in the North Pacific Ocean. The data are given as relative proportion (percentage) of each fragment within a profile, with the fragment size estimate (from the GeneScan program provided by ABI), rounded to the nearest integer. Based on this data set, we develop a DPA strategy for binning the profiles. The basic idea can be described as follows.

Each binning method can be regarded as a partitioning of the interval spanning of all fragment lengths. There are an exponential number of partitioning (binning methods) available and a criterion is needed to evaluate the quality of the binning. Intuitively, if the binning is appropriate, the replicate pairs, under the binning, should be similar in pairwise analysis. Therefore, our objective is to find a binning method minimizing the total distances between replicate pairs. We achieve this objective using dynamic programming.

With this objective, it is easy to notice that if we use one large bin for fragments of all sizes, the total distance would be zero. But this would have no practical value. Therefore, we impose a constraint on the maximum bin size such that the size of the bin ending at position i is at most MaxBinSize[i].

It has long been established that the likelihood as well as the range of sizing error increases as the ARISA/TRFLP fragment size increases (Fisher and Triplett, 1999; Kaplan and Kitts, 2003). Based on previous experimental data by Brown et al. (2005) and our preliminary examination of the profiles, we give the following maximum bin size, MaxBinSize, according to the length of the OTUs: MaxBinSize = 3 when fragment size is from 398 to 600 bp; 5 when fragment size is from 601 to 698 bp; 7 when fragment size is from 699 to 1200 bp. We let the MaxBinSize to be odd with the consideration that, for certain range of fragment sizes, the read from the machine can be up or down by (MaxBinSize[i] − 1)/2 bp.

**Data Preprocessing**

As is well known, systematic errors owing to human factors or equipment failure should first be identified and removed. Before the dynamic programming-based binning is applied, two data preprocessing steps were taken: (1) Removal of data with low relative abundance. With a cut-off of 0.5%, according to Hewson and Fuhrman (2004), only the relatively abundant OTUs are considered. (2) Removal of outliers. In general, the profiles of the replicate pairs of the same sample were very similar to each other. However, the profiles of a few replicate pairs look significantly different (Fig. 1).

Unlike univariate samples, for multivariate datasets the outliers no longer stand out at the ‘end’ of a sample (Barnett and Lewis, 1994). In order to detect outliers in this dataset, we compared profiles without any binning. In this case, qualitative match of position was more important than quantitative match of relative intensity. Therefore, the presence/absence data were used and the pairwise binary distances were computed for outlier detection. The basic hypothesis here is that the pairwise distances between replicates have the same distribution even if they are not based on the true binning. The distance for the outliers will be from another distribution with a large mean. Assuming that the pairwise distance follows a normal distribution, 95% of the sample is expected to fall into an interval centered around the mean with a spread of 1.96σ. With median instead of mean for robustness, we regarded the replicates outside this range as outliers.

**Dynamic programming based binning**

In this subsection, we describe the formulation of the problem, the criteria used to evaluate the quality of the binning method and the dynamic programming algorithm for finding the optimal binning.
Mathematical formulation If we think of a binning method as partitioning the fragment sizes spanning, say \([398, 1200]\), to a set of non-overlapping sub-intervals, each profile can then be re-coded as a new profile with data in each bin being merged according to the binning method. Our problem is the following: of all the possible partitions, we want to find the one (possibly not unique) that minimizes the overall sum of the distances between the recoded replicate profiles. The larger the number of bins, the more likely the binning method distinguishes different OTUs. Therefore, among the set of optimal solutions, we prefer the binning method with the largest number of bins.

Assume that all the fragments are between \([\text{Fstart}, \text{Fend}]\), where \(\text{Fstart} = 1\) and \(\text{Fend} = \text{Flen}\). The interval length is \(\text{Flen} = \text{Fend} - \text{Fstart} + 1\). In our case, \(\text{Fstart} = 398\), \(\text{Fend} = 1200\), \(\text{Flen} = 803\).

After data preprocessing, \(N = 42\) pairs of ARISA profiles taken at different time points at the same geographical location were available. The data can be regarded as \(42\) pairs of \(\text{Flen}\)-dimensional vectors with the \(l\)-th component being the intensity at fragment size \(397 + l\), \(l = 1, 2, \ldots, \text{Flen}\).

A binning method for the first \(S\) positions is represented by an increasing array and denoted as

\[
B^{S} = (B_{1}^{S}, B_{2}^{S}, \ldots, B_{k}^{S}, \ldots, B_{K(S)}^{S}),
\]

where each \(B_{k}^{S}\) is a bin boundary point; each bin, denoted by \(B_{k+1}^{S}\), is a half open interval: \((B_{k}^{S}, B_{k+1}^{S})\) and is of size no larger than \(\text{MaxBinSize}[B_{k+1}^{S}]\).

After a binning method is found, recoding is performed for each sequence.

DEFINITION 1. A recode of sequence \(R_{i}^{I}\) is the first sample of the \(i\)-th pair, with respect to binning method \(B^{S}\), is denoted as \(R_{i}^{I, B^{S}}\) and is constructed as follows: \(R_{i}^{I, B^{S}} = \left(\text{k}(S)\right)\) dimensional vector with the \(k\)-th element \(R_{i}^{I, B^{S}}k\) being the sum of observed intensity within the \(k\)-th bin, where \(k = 1, 2, \ldots, \text{k}(S)\).

DEFINITION 2. The quality score for the \(k\)-th bin \(B_{i}^{S}\), denoted by \(\text{Score}(B_{i}^{S})\), is defined as the Euclidean distance between the \(k\)-th elements from the recoded sequences of all the replicate pairs, i.e.

\[
\text{Score}(B_{i}^{S}) = \sum_{l=1}^{n_{1}} (R_{il}^{1} - B_{il}^{S})^2.
\]

DEFINITION 3. The quality score of the \(i\)-th replicate pair \(R_{i}^{1}\) and \(R_{i}^{2}\), with respect to a binning method \(B^{S}\), denoted by \(\text{Score}(R_{i}^{1, B^{S}}, R_{i}^{2, B^{S}})\), is defined as the Euclidean distance between \(R_{i}^{1, B^{S}}\) and \(R_{i}^{2, B^{S}}\), i.e.

\[
\text{Score}(R_{i}^{1, B^{S}}, R_{i}^{2, B^{S}}) = \sum_{l=1}^{n_{1}} (R_{il}^{1} - R_{il}^{2, B^{S}})^2.
\]

DEFINITION 4. The quality score for the binning method \(B^{S}\), denoted by \(\text{Score}(B^{S})\), is defined as the summation of the scores \(\text{Score}(R_{i}^{1, B^{S}}, R_{i}^{2, B^{S}})\) over all profile pairs, i.e.

\[
\text{Score}(B^{S}) = \sum_{i=1}^{N} \text{Score}(R_{i}^{1, B^{S}}, R_{i}^{2, B^{S}}).
\]

Or, the summation of the scores of all the bins, i.e.

\[
\text{Score}(B^{S}) = \sum_{k=1}^{K(S)} \text{Score}(B_{k}^{S}).
\]

Our objective is to find a binning method that minimizes the quality score \(\text{Score}(B^{S})\). This can be solved by dynamic programming. Among the binning method having the smallest quality score, the binning method having the largest number of bins is preferred.

Dynamic programming Suppose that \(B^{S}\) is the optimal binning method for the first \(S\) positions of the original data: \(R_{i}^{0}, R_{i}^{1}, i = 1, \ldots, N\) and \(s = 1, \ldots, S\). Let \(M(s) = \text{MaxBinSize}[s]\) be the maximal bin size of the bin that the \(s\)-th position belongs to. The minimal score for the first \(S\) positions is then the smallest score among the first \(S - i\) positions plus the score from the last bin of size \(i\), where \(1 \leq i \leq M(S)\).

Recursions

Based on the recursive idea, the minimal quality score for the first \(S\) positions is given by

\[
\text{Score}(B^{S}) = \min_{1 \leq i \leq M(S)} \left( \text{Score}(B^{S-i}) + \text{Score}\left(\left(S-i, S\right)\right) \right),
\]

and

\[
\{i_{\text{min}}\} = \arg \min_{i \leq M(S)} \left( \text{Score}(B^{S-i}) + \text{Score}(\left(S-i, S\right)) \right).
\]

The optimal binning method is given by

\[
B^{S} = B^{S-\text{min}} \cup \{S\}.
\]

DPA on presence/absence data As discussed above, there is some debate about the reliability of the fluorescence intensity as a measure of relative abundance, and presence/absence data are preferred in some studies. Therefore, we also modified our DPA to accommodate presence/absence data.

When DPA is applied to the presence/absence converted data, binary distance is used in the quality scoring function instead of the Euclidean distance.

For two sample profiles \(A\) and \(B\) with \(a_{i}\) and \(b_{i}\) denoting the presence/absence \((1/0)\) of each species \(i\), respectively. The binary distance of the two profiles is defined as

\[
\text{Dist}(A, B) = \frac{n_{00} + n_{11}}{n_{01} + n_{01} + n_{11}},
\]

where \(n_{kl} = \#\{i : a_{i} = k, b_{i} = l\}, k, l = 0, 1\).
Clustering analysis

With the binning obtained by DPA, the intensity data were binned. Hierarchical clustering with average linkage was applied on the binned data for the samples obtained at different times. The same clustering analysis was performed on the samples based on the major OTUs and the first 20 most abundant OTUs.

RESULTS

In this section, we present the results for data preprocessing, binning using DPA, and the clustering results based on the new binning method. We use both experimental data as well as examination of the profile data to determine the MaxBinSize vector as in the Materials and Methods section.

Data preprocessing

The binary distances based on the presence/absence profiles for the 47 replicate pairs are shown in Figure 2. By our data preprocessing procedure, 5 out of the 47 pairs of the ARISA data were found to be outliers. The replicate pairs for the 36th sample (12/16/2003) were exactly the same. Also, the Euclidean distance of the profiles for the replicate pairs 37 (2/19/2004), 38 (3/17/2004), 40 (5/19/2004) and 41 (6/17/2004) were too large. The five outliers were centered around the first half of 2004, indicating possible systematic errors. It was later found that during the time period of February–June 2004, the amount of DNA extracted from the bacteria was very small, and the outliers detected by our statistical approach coincide with the artifacts in data collection. The dataset with these five pairs removed (leaving 42 pairs of ARISA spectra data) were used for the following analysis. We also compared the binning results from the 42 samples with the binning results using only the the first 35 samples, as they are probably most likely reliable.

Dynamic programming based binning

We applied the DPA to the 42 pairs of ARISA data replicates with fragment lengths ranging from 398 to 1200 (803 sizes total). Using the percentage intensity data, a minimal quality score of 1400 was obtained resulting in 640 bins (including 122 non-empty bins and 518 bins containing no OTUs). The binning results for positions based on (397, 480] are given in Figure 3a. A bin is denoted by a half open interval with integer values. For example the first non-empty bin, (400, 402] contains fragment size of 401 and fragment size 402. It can also be seen that if the species with fragment length 401 was
recorded as present, the species with 402 was recorded as absent in most time spots. The most parsimonious explanation, considering the instrument precision, is that only the consensus species, in this case with fragment length 402 (present in the majority of samples), was correctly identified and the reported 401 fragment length was actually a 402 fragment length inaccurately sized. The consensus species here is defined as the species with the largest total sum of intensity over all the samples. Binning 401 and 402 together will reflect our best estimate about the abundance or presence/absence of the true OTU present. Many such examples can be listed, such as [418, 419, 420], [421, 422], [435, 436] in the part shown in Figure 3. The consensus species for the three bins are 419, 421 and 435, respectively. For the bin [435, 436], only 2 out of 42 replicate samples, (33, 37), show the existence of an OTU with fragment length 436 and both show the absence of fragment of size 435. It is reasonable to think that these two pairs are due to reading errors. There is a small percentage of both fragments 435 and 436 in replicate sample 37 which makes the DPA bin the two OTUs together.

Owing to the stochastic nature of PCR and potential systematic biases, the relative fluorescence intensity observed by ARISA may not be proportional to the true abundance in the community sample. However, profiles from duplicate samples show significant correlations between the observed intensities. The binning results based on intensity data are very informative for studying the diversity of microbial communities, permitting fair comparisons between samples while taking into account the accuracy and precision of the fingerprinting approach.

Although species with lower abundance may be as influential as species with high abundance, sometimes, an organism has importance to ecosystem function that is very disproportionate to its abundance. To consider this possibility, we also applied dynamic programming-based binning approach to the presence/absence form of the ARISA data, with binary distance as the quality score function. This resulted in a binning method with minimal binary distance of 0.12 and 629 bins with 117 non-empty bins, as shown in Figure 3b.

The binning results using the intensity data and the presence/absence data were highly correlated. Of the 117 right boundaries for the non-empty bins 97 coincide with the boundaries based on intensity data. There are 271 integers with at least one sample labeled as present. Assuming the 117 boundaries are randomly chosen from the 271 integers, the probability of observing at least 97 overlaps with the 122 right boundaries for the non-empty bins based on intensity data is

\[
\frac{\min (122, 117)}{(117 - i)} \frac{122}{i} \frac{271 - 122}{117 - i}.
\]

Therefore, the overlaps between the two partitions are significant with P-value close to 0 (2.97e–028).

Comparing Figure 3a with Figure 3b, we can see that the binary method based on presence/absence data tends to split bins based on intensity data when there are few pairwise mismatches. The primary reason is that binning based on intensity compares the sum of the intensity within a bin for the replicate pairs while binning based on presence/absence does not. As discussed above, [435, 436] were binned together based on intensity data. However, 435 and 436 were split into two bins based on presence/absence data. According to the above reasoning, 435 and 436 are most likely belonging to the same bin and the difference in reading probably reflects errors in calibration of fragment analysis, possibly from gel running conditions. Similarly, the bin [477, 478, 479] based on intensity data was split into two bins [477, 478] and [479] based on presence/absence data. [479] was present only in three replicate samples, with the simultaneous absence of bin [477, 478]. Most likely these readings are due to fragment analysis errors. Based on the above consideration, we prefer the binning results based on the relative intensity data.

The comparison between different binning methods is challenging because of the lack of existing criteria, and here we have tried to define such criteria. Previous studies have shown that the community profiles are closely related to the temperature. For each pair of samples, we calculated the absolute difference of their temperatures. The differences were sorted from the smallest and the ranks of the pairs of samples were obtained and denoted as \(R_t\). For a binning method B, we calculated the distance between each pair of samples based on the binned profiles. The order statistics based on the distances are denoted as \(R_B\). We calculated the correlation between \(R_t\) and \(R_B\) by permuting the temperatures of the 42 samples.

For the DPA-based algorithm, the correlation defined above is calculated as 0.14 with a P-value of 0.04, indicating the significant correlation between the temperature and the community profiles. For fixed sized binning, we set bin size at 3 for fragment size <500 bp, and 7 for fragment size >500 bp starting from fragment length 397–1200 bp. This strategy is used by Hewson et al. (2004). The correlation is calculated at 0.06 and the P-value is 0.318. Therefore, the DPA approach recovered more of the relationship between the temperature and the community profile.

Although samples 39 and 42–47 are not declared as outliers according to our criterion, they are probably not as reliable as the first 35 samples. Therefore, we also binned the profiles based on the first 35 samples. Based on intensity data, a total of 123 non-empty bins were found. Note that when 42 samples were used, 128 non-empty bins were found. The overlap between the two binning results is 100, which is statistically significant (P-value <5.46e–26).

**Clustering analysis and biological implications**

With the binning method from intensity data, the dataset was recorded and the mean of each replicate pair was taken as the profile for the time spot. The binning method gave 122 non-empty bins with some OTUs appearing only once or twice over the 4 year span, which indicates they are likely to be real OTUs but only occur occasionally or under very specific conditions. Therefore, we considered only the OTUs occurring at least a certain number of times—OTUs that occur at least three times in the 42 time spots. With this screening, 63 ‘major’ OTUs are left for the following clustering analysis.

Hierarchical clustering with average linkage was performed on the 42 time series samples. The dendrogram is shown in Figure 4a. The OTUs are ordered from the left to the right so that the OTUs with lower average intensity are on the left and the OTUs with higher average intensity are on the right. We further took the first 20 OTUs starting from the right as dominant OTUs and
performed clustering analysis on the time spots based on those dominant OTUs. The dendrogram is shown in Figure 4b.

As we can see, the two dendrograms clustered the samples in the same way. Therefore, we concentrate on Figure 4a. It is important to note the last cluster contains data after July 2004 with one exception (September 2003). There are two possible explanations. One is that the environment in the later half of 2004 was different from that in earlier years, making the microbial community significantly different from previous years. The other explanation is that systematic errors such as equipment/analysis error or human factors which made the profiles different. Note that Figure 2 gives the pairwise distances between duplicates for the different samples. Although the samples after July 2004 cannot be declared as outliers according to our criteria, the pairwise distances between replicates after July 2004 were still higher than those in previous years. Therefore, the most likely explanation is that these samples are possibly not as reliable as samples from previous years.

Next we concentrate on the data from August 2000 to December 2003. The hierarchical clustering algorithm divided the other samples into two main clusters. Cluster I contains 16 samples, mostly from August to January with three exceptions (March 2002, June 2001, June 2002). Cluster II can be further divided into two smaller clusters II-1 and II-2. II-1 contains three samples between August and October II-2 contains 16 samples mostly from March to July with three exceptions (December 2001, February 2001, September 2002). All the April and May samples belong to cluster II-2. In order to test the statistical significance of this observation, we calculate the probability that 16 randomly chosen samples (out of 36 samples, with all 2004 data excluded) contain at least 13 samples from March to July. Note that there are a total of 16 samples from March to July. Therefore, the probability is calculated as

$$\sum_{k=13}^{16} \binom{16}{k} \frac{\binom{36-16}{k}}{\binom{36}{16}} = 0.00009052$$

It is also interesting to see from the figure that the community has higher diversity during spring season. Further analysis on the biodiversity at different times will be studied in detail later.

**DISCUSSIONS**

In this paper, we developed a DPA-based method for binning microbial profiles obtained through ARISA. Prior to binning, we used statistical tools to identify outliers possibly owing to human or equipment errors. After the outliers were removed, we applied the DPA to the remaining data for profile binning which yielded reasonable binning results.

The maximum bin size for different fragment sizes were set based on previous experimental data as well as preliminary examination of the data. Experimental data for determining the dependence of error patterns on the fragment sizes are limited. The maximum bin sizes given in this paper are only an approximation. We also binned the data using other maximum bin sizes differing from those presented in this paper and the results were similar (data not shown). However, the maximum bin size vector cannot deviate too much from the true value or misleading results would occur. The MaxBinSize should not be too large. Otherwise different OTUs will be binned together making the binned results not distinguishing. Similarly, MaxBinSize should not be too small. Otherwise too much noise will be introduced into the binning results.

One might hypothesize that for the time series ARISA data at the same geographical location, the ARISA profiles in consecutive months are expected to be somewhat similar (i.e. the closer the months, the more similar the profiles). This information as well
as possible seasonality might be quite helpful in our binning. Currently the time information in the data is not used in the program, since the time spots are not continuous (some time spots are missing) and the time points are not exactly evenly distributed. But the binning results showing clustering of samples to some extent by season seem to reflect the time information. This may be used to describe the seasonal variation.

One of the limitations for this approach is that this algorithm considers only two replicates from the same sample. Sometimes more than two replicates may be available and all the replicate samples should be used in binning. To achieve this objective, a sound quality scoring function, the sum of the pairwise distances between any two replicates may be a good choice. With the current available experimental data, we cannot study the effect of more duplicates on the binning results. This is a topic of further study.

Another limitation is that currently only abundance higher than 0.5% is considered in this dynamic programming based binning approach. This will not significantly affect the OTUs with ‘occasionally’ low abundance but may miss some of the less abundant OTUs—the OTUs with consistently low abundance. How to balance errors with consideration of less abundant OTUs is an important problem.

We used a statistical method to identify samples in which the differences between the two duplicates are much higher than expected. The identified outliers are consistent with the fact that the DNA samples extracted during the identified time period are very small and thus the profiles for these samples may not be reliable. However, the statistical methods used in this study may fail to detect systematic errors due to equipment mal-function. In this situation, the two replicates may still have similar profiles because they are generated under the same mal-function condition. To detect such errors, more duplicates from each sample are needed.

In this paper, we refer to microbial species having 16S–23S rRNA intergenic spacer lengths within the same bin as an OTU and use the resulting OTUs as units of analysis. However, there are only a limited number of possible bins, but probably over several hundred thousand different bacterial species. Therefore, many different species may have the same 16S–23S rRNA intergenic spacer length and further binning of the intergenic spacer lengths may make each OTU contain many more species. While multiple distinguishable organisms occurring in one bin seems to be rare for the >0.5% ‘major’ OTU reported from our sampling location (Brown et al., 2005), in other environments it may be more of a problem. To understand the details of each OTU, more experiments using other markers, or the comparison of gene and ITS sequence differences are ideal.

In summary, a DPA for microbial profile binning was developed and applied to analyze a set of real ARISA profile data. The algorithm can also be used to other types of profiling/fingerprinting data such as profiles by DGGE, TRFLP, etc. The effects of maximum bin size and threshold for inclusion on the binning results needs to be studied further.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Ian Hewson in the Department of Ocean Sciences, University of California at Santa Cruz for his helpful comments about the background of the problem, Dr Ting Chen, in the Molecular and Computational Biology Program, University of Southern California, for his insightful suggestions and Huanying Ge in the Molecular and Computational Biology Program, University of Southern California, for his help in preparing the figures. The authors would like to thank the anonymous reviewers for their helpful comments. This research is supported by NSF grants MCB0084231 and OCE0527034.

Conflict of Interest: none declared.

REFERENCES


