BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 20 2006, pages 2532–2538
doi:10.1093/bioinformatics/btl417

Systems biology

Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors

Quansong Ruan[1], Debojyoti Dutta[2], Michael S. Schwalbach[3], Joshua A. Steele[3], Jed A. Fuhrman[3] and Fengzhu Sun[2,*]

[1]Department of Mathematics, University of Southern California, 3620 Vermont Avenue, KAP 108, Los Angeles, CA 90089-2532, USA, [2]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, MCB 201, Los Angeles, CA 90089-2910, USA and [3]Department of Biological Sciences, University of Southern California, 3616 Trousdale Pkwy, AHF 107, Los Angeles, CA 90089-0371, USA

Received on May 21, 2006; revised on July 22, 2006; accepted on July 26, 2006
Advance Access publication July 31, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Characterizing the diversity of microbial communities and understanding the environmental factors that influence community diversity are central tenets of microbial ecology. The development and application of cultivation independent molecular tools has allowed for rapid surveying of microbial community composition at unprecedented resolutions and frequencies. There is a growing need to discern robust patterns and relationships within these datasets which provide insight into microbial ecology. Pearson correlation coefficient (PCC) analysis is commonly used for identifying the linear relationship between two species, or species and environmental factors. However, this approach may not be able to capture more complex interactions which occur *in situ*; thus, alternative analyses were explored.

Results: In this paper we introduced local similarity analysis (LSA), which is a technique that can identify more complex dependence associations among species as well as associations between species and environmental factors without requiring significant data reduction. To illustrate its capability of identifying relationships that may not otherwise be identified by PCC, we first applied LSA to simulated data. We then applied LSA to a marine microbial observatory dataset and identified unique, significant associations that were not detected by PCC analysis. LSA results, combined with results from PCC analysis were used to construct a theoretical ecological network which allows for easy visualization of the most significant associations. Biological implications of the significant associations detected by LSA were discussed. We also identified additional applications where LSA would be beneficial.

Availability: The algorithms are implemented in Splus/R and they are available upon request from the corresponding author.

Contact: fsun@usc.edu

INTRODUCTION

Ribosomal RNA (rRNA) based whole community profiling techniques have become widely used in the study of microbial community composition. These approaches include denaturing gradient gel electrophoresis (DGGE) (Muyzer *et al*., 1993; Troussellier *et al*., 2002), terminal restriction fragment length polymorphism (TRFLP) (Avaniss-Aghajani *et al*., 1994; Liu *et al*., 1997; Saikaly *et al*., 2005) and automated rRNA intergenic spacer analysis (ARISA) (Fisher and Triplett, 1999; Ranjard *et al*., 2001; Brown *et al*., 2005). These methods are advantageous because they do not require cultivation or expensive sequencing and they take advantage of highly conserved rRNA genes (Pace *et al*., 1986). Molecular profiling of a community by these techniques generates highly reproducible community fingerprints, which allow ecologists to rapidly characterize the microbial community composition and the overall diversity for a large number of samples. One of the main goals of microbial ecologists is to identify temporal and spatial patterns in an effort to gain insight into microbial mediated ecological processes.

To better identify these patterns, several approaches have been employed in analyzing microbial community fingerprints, including principal component analysis (PCA), canonical correlation analysis (CCA), multidimensional scaling (MDS), discriminant function analysis (DFA), multiple linear regression and database examination in GenBank (Ranjard *et al*., 2001; Yannarell and Triplett, 2004, 2005; Fisher and Triplett, 1999; Stepanauskas *et al*., 2003; Van Mooy *et al*., 2004). Although many of these analyses yield interesting conclusions and insights, the techniques often require significant data reduction or data smoothing and can be difficult to extrapolate back to environmental factors. Alternatively, less elaborate techniques such as multiple linear regression or Pearson correlation analysis do not reduce large datasets to a handful of statistical variables, easing ecological interpretation, but often these techniques are too simplistic to characterize complex interactions observed in nature.

Moreover, most of these studies focused on the spatial and temporal variations of microbial community diversity rather than the variation dynamics among microbial species or the associations between microbial species and environmental variables, the latter of which may potentially lead to better understanding of the microbial and biogeochemical processes in marine ecosystems (Fuhrman *et al*., 2006).

[*]To whom correspondence should be addressed.

We believe that defining a robust similarity measure that exposes complex relationships between two species or between a species and an environmental factor is the first fundamental step in any rigorous analysis of microbial community fingerprints. In this paper, we address the problem of designing an appropriate similarity metric between temporal sequences of microbial species' abundance and the temporal sequences of the measurements of environmental factors. We also investigate 'co-varying' relationships among bacterial species and environmental factors. In this paper, we define a new graphical representation which enables us to study the co-varying relationships.

Pearson correlation coefficient (PCC) is commonly used to reveal relationships or associations. For two normally distributed random variables X and Y with observations $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, the PCC between x and y can be defined as

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y},$$

where s_x and s_y are the sample standard deviations of x and y, respectively, i.e.

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2} \text{ and } s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{y})^2}.$$

PCC is straightforward and powerful, and it captures the linear relationship between two bivariate normal distributed random variables. However, in reality, the above relationships can be much more complex, such as in gene regulation in biological systems, or in the community composition in ecological systems, where multiple species are related to different environmental factors in a highly non-linear fashion. Such complex relationships may not be captured by PCC analysis alone. For example, in gene expression analysis, the regulatory mechanism between ARG2 (acetylglutamate synthase) and CAR2 (ornithine aminotransferase) depends on the expression level of CPA2 (arginine specific carbamoylphosphate synthase), but the PCC between the expression levels of ARG2 and CAR2 is nearly zero (Li, 2002).

Several efforts have been made to resolve such issues in time-course gene expression data analysis. Similar to the local alignment algorithm in DNA sequence alignment (Waterman, 1995), Qian *et al*. (2001) proposed a local clustering approach to reveal the various types of relationships between the yeast gene expression profiles. Similar idea has been used based on local shape similarity to analyze gene expression profiles (Balasubramaniyan *et al*., 2005).

The inherent difficulties in cultivating marine bacterioplankton may suggest that bacteria live in highly connected communities. Therefore, the presence and abundance of individual taxa within the bacterioplankton are likely to be dependent on each other as well as on their surrounding environment. Moreover, these relationships are likely to be nonlinear due to limiting resources and fluctuating environments. Thus, we also need to explore alternative similarity metrics.

In this paper, we propose a Local Similarity metric to reveal such associations among bacterial taxa [here defined simply as operational taxonomic units (OTUs) (In this paper, each OTU is represented by an index number for convenience. The corresponding fragment length can be looked up from Table 3 in the Appendix section.)] and with environmental factors. In order to study complex associations between OTUs and the environment, we construct a novel graphical representation that depicts the variations in OTU abundance relative to the environmental factors as a directed, annotated, weighted graph, which we call the co-varying graph. We demonstrate, using the Actinobacteria subgroup as an example, that this graph makes it much easier to reveal the potential ecological niche of individual taxa. We note that our similarity metric in conjunction with the graphical representation could lead to new algorithmic methods to analyze microbial community fingerprints.

This paper is organized as follows. First, we described the local similarity analysis (LSA) in the Materials and Methods Section. Second, we demonstrated the efficacy of LSA using simulated data, after which we applied LSA to a marine microbial observatory dataset to identify significant associations among marine bacterioplankton as well as significant associations among the marine bacterioplankton and the environmental factors. Finally, we discussed some variations or extensions of LSA and other possible applications, followed by some limitations of this approach in the Discussion section.

MATERIALS AND METHODS

Materials

Our study was based on a marine bacterioplankton community ARISA time series sampled between August. 2000 and December. 2004, approximately once per month from the same geographical location in San Pedro Channel in the North Pacific Ocean. ARISA was performed on each sample, which provides data for our analysis. Each data point of the time series was given as a profile with relative percentage of each fragment size estimate within the profile and measurements of 14 environmental factors: virus counts, bacterial counts, tdr counts, leu count, salinity, oxygen density, nitrite (NO_2) density, nitrate (NO_3) density, phosphate (PO_4), chlorophyll a concentrations, Phaeopigment concentrations, temperature and stratification index (SigmaTheta).

There were missing data in our dataset. In particular, entries of measurements of some environmental factors were missing. Several approaches are available for data imputation. Since the time points in our dataset were only one month apart, we assumed gradual, linear rates of change in environmental factors occurred. Therefore, for both simplicity and reality, we used linear interpolation to fill in the missing data entries.

The outlier detection procedure (Ruan *et al*., 2006) found some outliers in the sample, mainly from the year 2004, which indicates that data from year 2004 might not be as reliable. Therefore, only the data from 2000 to 2003 were used in our analysis.

Furthermore, the original ARISA dataset contains imprecise fragment sizes from sources such as sequencing and gel electrophoresis. These fragment sizes need to be converted to their appropriate 'bins' for the following community composition variation analysis. Currently several fixed size binning approaches (Stepanauskas *et al*., 2003; Hewson and Fuhrman, 2004) and variable size binning approach (Ruan *et al*., 2006) have been proposed for this purpose. Here, we took the dynamic programming based variable size binning because this approach allows variable bin sizes to minimize the overall difference between replicate profiles and is based on data itself rather than a priori assumptions about technological errors (Ruan *et al*., 2006).

After binning, the final dataset contained 58 major OTUs (OTUs which were present at three or more time points) with their corresponding relative abundance, 14 environmental factors and 35 time points for analysis.

Methods

Normal transformation For two variables X and Y with observations $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ that are not normally distributed, before their PCC and local similarity are computed, normal score transformation (Li, 2002) is performed for the two sequences. For variable x, suppose its rank vector is $R^x = (R^x_1, R^x_2, \ldots, R^x_N)$, then the normal transformed sequence is $x' = (x'_1, x'_2, \ldots, x'_N)$, where $x'_i = \phi_{-1}(\frac{R^x_i}{N+1})$, for $i = 1, \ldots, N$. PCC of x' and y' is computed from the normal score transformed sequences, x' and y'. The statistical significance of $r(x, y)$ is tested by the fact that $t = r\sqrt{\frac{N-2}{1-r^2}}$ has t-distribution (degree of freedom: $\upsilon = \mathrm{N} - 2$ with mean 0 and variance $v/(v-2)$. The local similarity score (as defined in the following section) between x and y is also computed from the normal transformed sequences.

Local similarity For two normal transformed sequences of the same length, the local similarity score is defined as the maximal sum of the product of the corresponding entries of all their subsequences within some predefined time delay D. The value of D indicates how far in time the species interaction can be. We choose $D = 3$ in our analysis. The local similarity score is computed by dynamic programming.

Local similarity between two OTU time series: Suppose the two OTU normal transformed series are O_1 and O_2 with both of length n, i.e. $O_1 = \mathrm{O}_{11}, O_{12}, \ldots, O_{1n}$ and $O_2 = O_{21}, O_{22}, \ldots, O_{2n}$. The positive score matrix $P_{n \times n}$ and negative score matrix $N_{n \times n}$ are calculated as follows:

(1) For $i, j = 1, \ldots, n$,
$P_{0,i} = P_{j,0} = 0$ and $N_{0,i} = N_{j,0} = 0$.

(2) For $i, j = 1, \ldots, n$ with $|i - j| \leq D$,
$P_{i+1,j+1} = \max\{0, P_{i,j} + O_{1,i+1} \cdot O_{2,j+1}\}$ and
$N_{i+1,j+1} = \max\{0, N_{i,j} - O_{1,i+1} \cdot O_{2,j+1}\}$.

(3) $P(O_1, O_2) = \max\limits_{1 \leq i,j \leq n} P_{i,j}$ and
$N(O_1, O_2) = \max\limits_{1 \leq i,j \leq n} N_{i,j}$.

(4) $\mathrm{MaxScore}(O_1,O_2) = \max(P(O_1,O_2), N(O_1,O_2))$ and
$\mathrm{Flag}(O_1,O_2) = \mathrm{sgn}(P(O_1,O_2) - N(O_1,O_2))$.

The local similarity score of two time series O_1 and O_2, $LS(O_1,O_2)$ is defined as the $\mathrm{MaxScore}(O_1,O_2)$ divided by the length of the time series n, i.e.

$$LS(O_1, O_2) = \frac{\mathrm{MaxScore}(O_1, O_2)}{n},$$

whether the score is for positive or negative correlation of the two sequences is given by the sign function $\mathrm{Flag}(O_1,O_2)$.

Complexity: For two sequences of length n, in general, such dynamic programming based algorithm takes $O(n^2)$ time. But in our approach a constraint D (constant and in general small compared to n) is imposed on the delay of the two matching subsequences, the complexity reduces to $O(n)$. For pairwise local similarity scores of m such sequences, there are $O(m^2)$ such local similarity scores to compute; therefore, the total complexity is $O(m^2n)$.

Local similarity between OTU and environmental factors: The local similarity score between an OTU and an environmental factor is defined similarly to local similarity scores between OTUs. The difference is that for some environmental variables such as virus counts, bacterial counts, tdr, and leu are log-scaled before normalization is performed for each environmental factor series.

Statistical significance: In order to see whether the local similarity scores are statistically significant or just obtained by chance, the statistical significance level represented by P-value is computed. This was performed by permutation test: (1) For each pair of OTU sequences O_1 and O_2 with some local similarity score $LS(O_1,O_2)$, we first randomly permute the values of the data in each sequence. (2) Then local similarity score is computed for the permuted sequences following the above procedure. (3) Repeat step (1) and (2) for a large number of times (say, $N = 1000$), an empirical distribution of the local similarity score is generated. (4) Compute the probability that the score is as high as $LS(O_1,O_2)$. In our case simply compute the proportion of times that the scores are at least as high as the score $LS(O_1,O_2)$ and assign this proportion as the P-value for the local similarity score of O_1 and O_2.

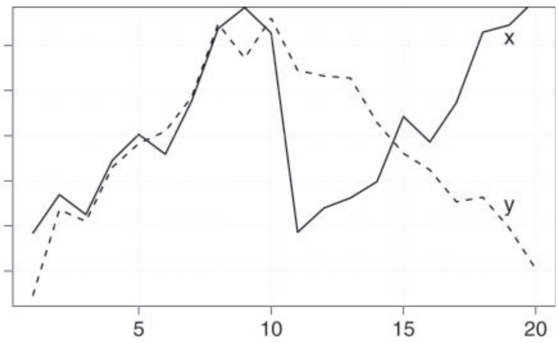

Fig. 1. Local correlations identified by local similarity but not by Pearson correlation. A total of 20 observations of two random variables X and Y are simulated. The first 10 observations are generated by $Y = X + \epsilon$ and the latter 10 observations are generated by $Y = -X + \epsilon$, where ϵ is standard normal, i.e. $\epsilon \sim N(0,1)$.

Note that $\binom{58}{2} = 1653$ OTU pairs and $58 \times 14 = 812$ OTU-environmental factor pairs are studied. Multiple testing issues need to be considered. Bonferroni correction is too conservative in this study. Instead, we use the false discovery rate (FDR) or Q-value to adjust for multiple testing. Q-value is defined as the fraction of false positives if a given pair is declared as significant. We use the software QVALUE (Storey and Tibshirani, 2003) (http://faculty.washington.edu/jstorey/qvalue/) to do our analysis.

Co-varying graph After the statistically significant OTU pairs or the OTU-factor pairs have been obtained by a similarity metric, we now construct a novel graphical representation of the relationships. We define a graph $G(V, E)$, where V is the node set and E the edge set. For any $v \in V$, v could be either an OTU or an environmental factor. For any edge $e(u,v)$, there is a relationship between u and v of type $t(u,v)$, where t is either undirected or directed, with an arrow pointing towards v if it is delayed in comparison to u. Edge $t(u, v)$ is solid if the local similarity score is positive and dashed when the score is negative. An OTU node is represented by a circle and an environmental factor node is represented by a rectangle.

For visualization purpose, the co-varying graph is drawn by Cytoscape Version 2.2 (Shannon *et al*., 2003).

RESULTS

Simulations

In this section, we generate simulated data and show the efficacy of our local similarity metric.

Local Correlations A total of 20 simulated observations of two random variables X and Y are shown in Figure 1. Here we assume that $Y = X + \epsilon$ for the first 10 observations and $Y = -X + \epsilon$ for the latter 10 observations, where ϵ is standard normal, i.e. $\epsilon \sim N(0,1)$.

The PCC for the two variables is −0.126. How to interpret this statistic is unclear because visual inspection of Figure 1 suggests a nonlinear relationship between the variables x and y. Indeed if we look further into these two variables, we notice that there are strong positive association between x and y in the first half of the sequences and negative association in the second half of the sequences (the Pearson correlation coefficients 0.91 and

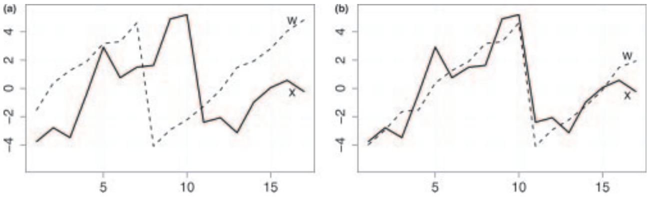

Fig. 2. Correlations with time delay identified by local similarity but not by Pearson correlation. Two temporal sequences with some time delay were simulated by $W_{n+3} = X_n + \epsilon$, where ϵ is standard normal, i.e. $\epsilon \sim N(0,1)$. (**a**) Original sequences. (**b**) Aligned view.

−0.94, respectively), LSA catches this association and the normalized local similarity score is 0.475 with a *P*-value of 0.044.

Correlations with time delay Two temporal sequences with some time delay were simulated (Fig. 2). Here we assume that $W_{n+3} = X_n + \epsilon$, where ϵ is standard normal, i.e. $\epsilon \sim N(0, 1)$. The PCC between the two sequences is −0.122 (*P*-value 0.32).

As we can see from Figure 2a that there is a three time points delay between the two temporal sequences—sequence *x* is three time points behind sequence *w*. The aligned view of the two sequences show a strong correlation, as can be seen in Figure 2b. LSA can identify this time delay—it gives a local similarity score of 0.653 (with a *P*-value of 0.0273) between these two sequences . Of course, if we knew this time delay beforehand, we could compute the Pearson correlation of the two matching subsequences.

Marine bacterioplankton community profile analysis

We applied LSA to a real ecological dataset which included measurements of bacterial community composition via ARISA as well as 14 environmental factors sampled at 35 time points as mentioned in the Materials and Methods section. The dynamic program algorithm based binning method (Ruan *et al.*, 2006) defined 58 major OTUs. The 58 OTUs and 14 environmental factors form the basis of our analysis.

LSA revealed 249 OTU pairs and 128 pairs of OTUs and environmental factors with *P*-value ≤0.05. With more stringent significance level of 0.01, 100 OTU pairs and 44 pairs of OTU and environmental factors with significant local similarity scores were found.

The histogram for the resulting *P*-values for all the OTU pairs is given in Figure 3a. In order to see the expected distribution of the *P*-values under the null model of no association among the OTUs, we randomly permute each OTU sequence to obtain a total of 58 permuted OTU sequences. The local similarity scores and the corresponding *P*-values are calculated similarly as above. Figure 3b gives the histogram for the *P*-values for the permuted data which closely resembles a uniform distribution in [0,1]. Figure 3a shows that the *P*-values for the real data are biased toward 0 indicating many closely related OTU pairs.

Furthermore, as we computed the PCC for the OTU pairs and pairs of OTUs and environmental factors with significant local similarity scores, we also found that 186 out of the 249 significant OTU pairs by local similarity are also significant by PCC (*P*-value 0.05).

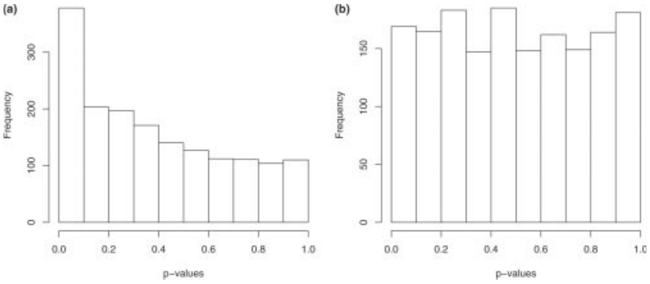

Fig. 3. The histograms of the *P*-values for (**a**) the real data and (**b**) the permuted data.

Table 1. OTU paris with *P*-value ≤0.01 using LSA, but not using PCC

OtuX	OtuY	LS score(*P*-value, *P*-value)	PCC(*P*-value)
10	31	−0.442(0.007, 0.09413)	−0.172(0.158)
17	31	−0.411(0.007, 0.09413)	−0.212(0.107)
22	31	+0.399(0.004, 0.06901)	0.353(0.017)
30	56	−0.382(0.005, 0.07749)	−0.212(0.108)
11	36	+0.360(0.006, 0.08708)	−0.059(0.367)
15	48	−0.342(0.009, 0.10757)	−0.283(0.047)
2	7	+0.330(0.001, 0.03208)	0.116(0.249)
5	7	−0.329(0.007, 0.09413)	0.262(0.061)
5	54	−0.303(0.009, 0.10757)	−0.083(0.315)
30	58	+0.293(0.009, 0.10757)	−0.092(0.297)
1	54	−0.288(0.008, 0.10679)	−0.283(0.047)
19	34	+0.280(0.010, 0.11289)	−0.273(0.054)
15	34	+0.263(0.009, 0.10757)	0.130(0.225)
41	54	+0.259(0.006, 0.08708)	0.168(0.164)
13	34	+0.252(0.001, 0.03208)	−0.033(0.424)
5	43	−0.248(0.002, 0.04572)	0.175(0.153)
28	57	−0.229(0.009, 0.10757)	−0.248(0.072)
27	57	+0.226(0.006, 0.08708)	0.318(0.029)

The first two columns are the OTU pairs. The third column gives the LSA scores, the *P*-values and the *Q*-values. The fourth column shows the PCC score and the corresponding *P*-value.

LSA also revealed significant relationships (*P*-value ≤ 0.01) among OTUs as well as OTU and environmental factors which could not be detected using PCC. Specifically, of the 100 significant OTU–OTU associations detected by LSA, 18 associations were found to be not significant by PCC and 11 out of 44 significant association pairs of OTU and environmental factors are non-significant by PCC. For each *P*-value of the local similarity score, a *Q*-value was calculated.

Co-varying OTU pairs identified by local similarity but not by PCC Out of the 100 significant OTU–OTU associations, 18 associations were found to be not significant by PCC (*P*-value ≤ 0.01) (Table 1).

OTU pair (22,31): Figure 4. This OTU pair OTU22 (fragment length: 658) and OTU31 (fragment length: 718) are putative α Proteobacteria within the SAR11 group. They showed high local similarity score of 0.399 with *P*-value of 0.004. The *Q*-value of 0.06901 indicates that for this significant OTU pair, there is a probability of ∼0.07 that it is a false discovery, i.e. there is indeed

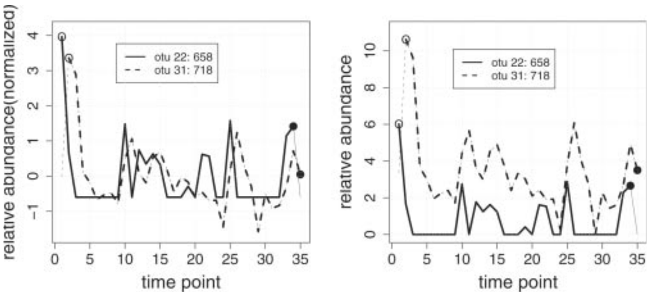

Fig. 4. Significant OTU pair: OTU 22 (fragment length: 658) and OTU 31(fragment length: 718). Open and closed circles. indicate the starting point and ending point of matching subsequences. Left panel: normalized plot. Right panel: relative abundance as percentage.

Table 2. Examples of OTU and environmental factor pairs with significant local similarity scores but non-significant PCC

OUT	Factor	LS score(*P*-value, *Q*-value)	PCC(*P*-value)
48	63(sal)	+0.501(0.001, 0.03208)	0.380(0.011)
25	69(chla)	−0.471(0.002, 0.04572)	−0.323(0.027)
48	59(virus)	−0.436(0.004, 0.06901)	−0.318(0.029)
2	62(leu)	−0.423(0.005, 0.07749)	−0.207(0.113)
48	64(oxy)	−0.419(0.010, 0.11289)	−0.206(0.114)
7	71(temp)	+0.375(0.002, 0.04572)	−0.088(0.305)
50	66(no3)	+0.365(0.006, 0.08708)	−0.071(0.340)
7	72(sigma)	−0.342(0.002, 0.04572)	0.059(0.365)
32	66(no3)	+0.329(0.002, 0.04572)	−0.055(0.375)
54	71(temp)	−0.315(0.003, 0.06096)	−0.188(0.136)
44	67(sio3)	+0.240(0.003, 0.06096)	0.214(0.105)

no association between this pair. This OTU pair also showed an interesting one time point delay with OTU22 being ahead of OTU31. The similarity pattern is clear in Figure 4b. Although only direct experimentation can determine the mechanism underlying the observed patterns, these results may suggest a delayed association between OTU22 and OTU31.

We can see from their pairwise plots that these OTUs track each other closely albeit with a time delay of one month, which may explain why PCC was not able to detect a direct linear relationship.

OTU and environmental factor pairs with significant local similarity scores but non-significant PCC Pairs of OTU and environmental factor with significant local similarity scores but non-significant PCC were also found, as shown in Table 2.

Interactions between Actinobacteria and bacterial production: In the significant pairs, we found that OTU2 (fragment length: 420), which is putatively identified as an Actinobacteria, has significant local similarity score of 0.423 with ENV62(leu) with *P*-value of 0.005 and *Q*-value of 0.007749. While the PCC of this OTU pair is −0.207 (Fig. 5).

As can be seen from Figure 5, this OTU exhibits a delayed but significant negative correlation with bacterial production as measured by the uptake rate of tritiated leucine (leu), possibly suggesting a life history strategy uniquely adapted to low energy,

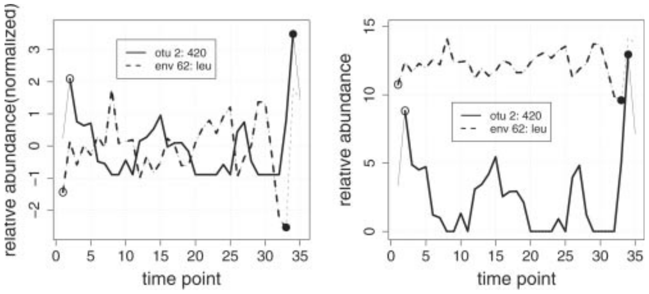

Fig. 5. Significant pair: OTU 2 (fragment length: 420) and ENV62(leu). Open and closed circles indicate the starting point and ending point of matching subsequences. Left panel: Normalized plot. Right panel: relative abundance as percentage.

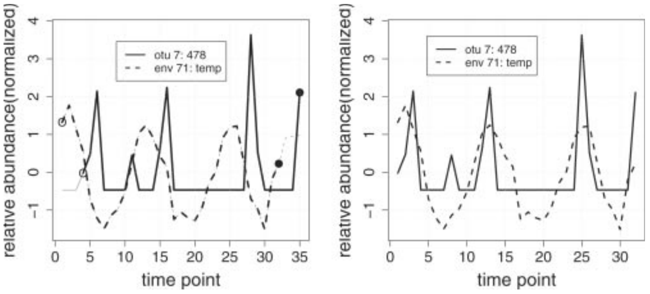

Fig. 6. Significant pair: OTU7 (fragment length: 478) and ENV71 (temperature). Open and closed circles indicate the starting point and ending point of matching subsequences. Left panel: normalized plot. Right panel: aligned view.

oligotrophic ocean gyres. This putative correlation can be identified by LSA but not by PCC. Moreover, the probability that this discovery is wrong (i.e. OTU2 and ENV62 are indeed not correlated) is ~0.08, indicated by its *Q*-value.

*Interaction between SAR*86 *and temperature*: OTU7(478) versus ENV71(temp): Figure 6. In this figure, OTU 7(fragment length: 478) is a putative member of the yet uncultivated SAR86 subgroup of γ-Proteobacteria. Figure 6 illustrates a clear but delayed covariance between OTU 478 abundance and temperature. The time delay is larger in this instance, however, with OTU7(478) reaching its highest abundance ~3 months after the temperature reaches its highest point each year. This result illustrates the importance of long term observatories in ecological studies. Understanding why this OTU so closely tracks temperature will likely be the focus of future analysis and experimentation. Given the peak temperature is strongly seasonal, it is likely that this OTU prefers overall conditions that occur during autumn, in agreement with our recent observation that many OTU have strong seasonally repeating patterns (Fuhrman *et al*., 2006).

A co-varying graph

Based on the results from the LSA and results from PCC, a co-varying graph can be constructed. As an example, we constructed a co-varying graph for four Actinobacteria subgroups: OTU 420 (Actinobacteria group I), OTU 422 (Actinobacteria group II), OTU 424 (Actinobacteria group III) and OTU 436 (Actinobacteria group IV).

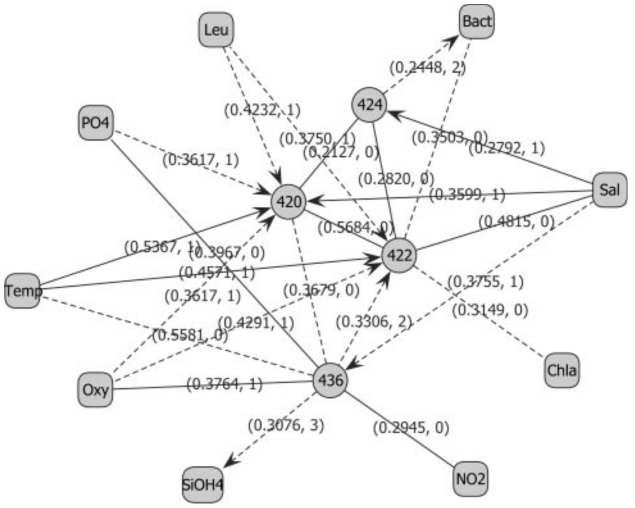

Fig. 7. Co-varying graph for four Actinobacteria subgroups with environmental factors. Constructed based on the significant local similarity scores between OTUs or scores between OTUs and environmental factors (*P*-value 0.05). Each node represents an OTU (circle) or an environmental factor (square). Solid edge represents positive association, dashed edge represents negative association. Directed edge indicates time delay between two OTUs or an OTU and an environmental factor with arrow pointing to the OTU or factor that is behind the other in time. Each edge is labeled by the association score followed by the time shift.

Among the 4 Actinobacteria subgroups, 25 OTU pairs or OTU and environmental factor pairs were found with significant local similarity scores (P-value ≤ 0.05).

In the co-varying graph, each OTU is represented by a circle denoted by the fragment length of the OTU. Each environmental factor is represented by a square box denoted by its name. If there is positive association between an OTU and another OTU (or an environmental factor), a solid edge is created between them; otherwise (negative association), a dashed edge is created. If there is time delay between an OTU pair or an OTU and an environmental factor, the edge is a directed edge pointing to the node that is behind the other in time. Each edge is labeled by the association score followed by the time shift. According to the above criteria, the resulting co-varying graph for Actinobacteria subgroup is shown in Figure 7.

From the co-varying graph (Fig. 7), we can see that three of the Actinobacteria subgroups are strongly associated with environmental factors, including Actinobacteria subgroup I (OTU420), Actinobacteria subgroup II (OTU422) and Actinobacteria subgroup IV (OTU436). These three subgroups show consistent associations with temperature (Temp), oxygen (Oxy) and salinity (Sal). Actinobacteria subgroup III (OTU424) have significant associations with only salinity (Sal) and bacterial counts (Bact), indicating that this subgroup might prefer different environmental conditions from the other three. Out of the three subgroups, only Actinobacteria subgroup II (OTU422) shows significant negative association with chlorophyll a concentrations (Chla) and only Actinobacteria subgroup IV (OTU436) showed slight association with nitrite (NO_2). We can also observe from the graph that oxygen (Oxy) was negatively tracked by two Actinobacteria subgroups I and II (OTU420 and OTU422) by one time point. Out of the environmental factors, only one factor salinity (Sal) showed significant association with all of the four subgroups. Furthermore, it was tracked closely by three subgroups I, III and IV (OTU420, OTU424 and OTU436) by one time point. LSA results indicate that the subgroups of the Actinobacteria clade identified by ARISA do not all respond to environmental factors in an identical manner. This suggests that the Actinobacteria subgroups may represent distinct ecotypes occupying different niches at the site studied.

DISCUSSION

LSA is a computational approach that is used to extract relationship among variables of interests and works well with time series datasets. It is capable of extracting the general linear regression statistic as well as identifying co-variance relationships potentially masked by time delays or associations with external dependence (although it does not identify the external dependence). Similar dynamic programming based approaches have been applied for DNA sequences alignment (Waterman, 1995) and gene expression analysis (Qian *et al.*, 2001).

From a multi-year time series profile, Brown *et al.* (2005) obtained a seasonal profile of each taxon by averaging the profiles from the same month and identified some temporal patterns of the OTUs. We could perform our analysis based on the collapsed time series as well. But since the biochemical/environmental conditions might not be the same in the same month of different years, averaging of the multi-year time series to a 12 month sequence might result in too much smoothing of the natural variations inherent within the dataset.

The co-varying graph offers an intuitive way for visualizing the interactions between species and with environmental factors. Based on the results from LSA, the co-varying graph shows that the abundance of OTUs are associated with each other as well as with the environmental factors. The relationships from this analysis are putative and need further biological experiments or other ways for verification.

In this paper, local similarity was introduced for analyzing relationships between bacterial species more as a general concept. Some variation/extensions are possible.

Currently, LSA has been applied to the time series of the relative abundance of the OTUs in the community. In a situation where the contribution of the OTUs is to be considered equally regardless of abundance, LSA can be adjusted to work with a presence–absence matrix. In this case, our data matrix needs to be converted to a 0–1 matrix, where 0 denotes absence and 1 denotes presence, with a certain detection threshold. Then the corresponding local similarity algorithm can be easily developed which will be more similar to the local alignment for DNA sequence analysis.

The application of LSA directly on the relative abundance data (including the presence/absence case) is focused on the ‘coexistence’ of OTUs. This means, as long as two OTUs are present in the community at the same time, they usually have a high similarity score. But this high score does not fully uncover the ‘co-varying’ relationship between OTUs. Therefore, an OTU abundance data matrix only gives us the relative abundance of the OTUs in the community. If we are more interested in the ‘co-variance’ of the OTUs temporally, rather than only the ‘coexistence’ of the OTUs, as we considered above, it might be more appropriate to look at their temporal changes, i.e. the changes of the OTU relative abundance in the community between consecutive time points. For example, OTU *i* might have abundance of 0.10 in May but 0.12 one month later—June, then the relative change of abundance from May to June is $(0.12 - 0.10/0.10) = 0.20$. For an abundance time series of length L, we will obtain a relative change sequence of length L-1. As an extension, we can apply LSA on the relative change sequences of the original time series.

Moreover, in order to illustrate its capability, LSA was only applied to an ARISA community profiling time series. This approach can also be used for analyzing community profiling time series datasets obtained from other molecular fingerprinting techniques such as DGGE and TRFLP as well as any species-abundance data from other non-molecular techniques. Indeed, this approach is not limited to microbial community analysis. It is also possible to apply it on species-abundance data for macro-organisms such as plants or animals.

Our results suggest that LSA is an effective statistical analysis approach for rapidly detecting numerous associations among a large number of potentially related variables. The results from this study lend insight into marine microbial ecology as well as to facilitate the design of future experimental studies.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Ting Chen in the Molecular and Computational Program at the University of Southern California (USC) for his helpful comments and support. The authors also thank the Microbial Observatory at the University of Southern California (USC), the National Science Foundation Microbial Observatories program, the captain and crew of the R/V Sea Watch, Reni Smith, Michael Neumann, Ian Hewson, Sheila O'Brien, Anand Patel, Troy Gunderson, Pete Countway, Adriane Jones, Ilana Gilg for collecting the samples and preparing the profiling data. The authors thank two anonymous reviewers for their helpful comments. This work was supported by NSF Grants MCB0084231 and OCE0527034.

Conflict of Interest: none declared.

REFERENCES

Avaniss-Aghajani,E. *et al*. (1994) A molecular technique for identification of bacteria using small subunit ribosomal RNA sequences. *Biotechniques*, **17**, 144.6–148.9.

Balasubramaniyan,R. *et al*. (2005) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, **21**, 1069–1077.

Brown,M.V. *et al*. (2005) Coupling 16S-ITS rDNA clone libraries and ARISA to show marine microbial diversity; development and application to a time series. *Environ. Microbiol*., **7**, 1466–1479.

Fisher,M.M. and Triplett,E.W. (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl. Environ. Microbiol*., **65**, 4630–4636.

Fuhrman,J.A. *et al*. (2006) Annually reoccurring bacterioplankton communities are predictable from ocean conditions. *Proc. Natl Acad. Sci. USA*, in press.

Hewson,I. and Fuhrman,J.A. (2004) Richness and density of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl. Environ. Microbiol*., **70**, 3425–3433.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad. Sci. USA*, **99**, 16875–16880.

Liu,W.T. *et al*. (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl. Environ. Microbiol*., **63**, 4516–4522.

Muyzer,G. *et al*. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction--amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol*., **59**, 695–700.

Pace,N.R. *et al*. (1986) The analysis of natural microbial populations by ribosomal RNA sequences. *Adv. Microbial Ecol*., **9**, 1–55.

Qian,J. *et al*. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol*., **314**, 1053–1066.

Ranjard,L. *et al*. (2005) Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. *Appl. Environ. Microbiol*., **67**, 4479–4487.

Ruan,Q. *et al*. (2006) A dynamic programming algorithm for binning microbial community profiles. *Bioinformatics*, **22**, 1508–1514.

Saikaly,P.E. *et al*. (2005) Use of 16S rRNA gene terminal restriction fragment analysis to assess the impact of solids retention time on the bacterial diversity of activated sludge. *Appl. Environ Microbiol*., **71**, 5814–5822.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Stepanauskas,R. *et al*. (2003) Covariance of bacterioplankton composition and environmental variables in a temperate delta system. *Aquat. Microb. Ecol*., **31**, 85–98.

Sorey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide experiments. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Troussellier,M. *et al*. (2002) Bacterial activity and genetic richness along an estuarine gradient (Rhone River plume, France). *Aquat. Microb. Ecol*., **28**, 13–24.

Van Mooy,B.A. *et al*. (2004) Relationship between bacterial community structure, light, and carbon cycling in the eastern subarctic North Pacific. *Limnol. Oceanogr*., **49**, *282–288*.

Waterman,M.S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall/CRC, NY, USA.

Yannarell,A.C. and Triplett,E.W. (2004) Within- and between-lake variablility in the composition of bacterioplankton communities: investigations using multiple spatial scales. *Appl. Environ. Microbiol*., **70**, 214–223.

Yannarell,A.C. and Triplett,E.W. (2005) Geographic and environmental sources of variation in lake bacterial community composition. *Appl. Environ. Microbiol*., **71**, 227–239.

APPENDIX

OTU mapping

For convenience, OTUs are represented by their indices instead of their actual fragment lengths followed by environmental factors represented by their names in short. Following is the OTU/env factor mapping Table A1.

Table A1. Index mapping of the 58 major OTUs and 14 environmental factors

	0	1	2	3	4	5	6	7	8	9
0	—	402	420	422	424	436	473	478	489	521
1	538	561	573	582	593	617	619	622	628	631
2	647	656	658	662	669	677	683	692	704	707
3	711	718	723	728	739	741	749	755	757	761
4	768	772	775	779	787	791	798	806	828	839
5	855	858	887	908	967	986	1051	1129	1186	virus
6	bact	tdr	leu	sal	oxy	no2	no3	sio3	po4	chla
7	phaeo	temp	sigma							

Example: OTU with index 25 corresponds to fragment length of 677, which is located at the third row (with row number of 2) and the sixth column (with column number of 5). The name of environmental factor with index 59 is located at 6th row(with row number 5) and the 10th column (with column number 9), which is environmental factor 'virus'.