Action: respond to our copy-editing questions

Select each question and describe any changes we should make on the proof. Changes against journal style will not be made and proofs will not be sent back for further editing.

AQ1: Please check the spelling and accuracy of all author names and affiliations, particularly for any of your co-authors. Please also check that author surnames are correctly identified with a pink background. This is to ensure that forenames and surnames are tagged correctly for online indexing. **Incorrect names and affiliations may lead to an author not being credited for their work by funders, institutions, or other third parties.**

AQ2: If your manuscript has figures or text from other sources, please ensure you have permission from the copyright holder. For any questions about permissions contact jnls.author.support@oup.com.

AQ3: Please check that funding is recorded in a separate funding section if applicable. Use the full official names of any funding bodies, and include any grant numbers.

AQ4: Journal policy requires authors to provide a data availability statement in their manuscript. Please confirm that this statement is included in your manuscript and that any required links or identifiers for your data are present in the manuscript as described or provide edits with the required information.

AQ5: AU: Please provide the FAX of the corresponding author.

AQ6: AU: Please define all abbreviations not well known to readers at first mention in the abstract and text and also in the article title.

AQ7: The text/caption for Figure 2 currently contains reference to colour used within the figure. For accessibility purposes reference to colour should be avoided. If possible please provide updated wording for the figure caption to remove reference to colour.

AQ8: AU: Please provide a Funding statement, detailing any funding received. Remember that any funding used while completing this work should be highlighted in a separate Funding section. Please ensure that you use the full official name of the funding body, and if your paper has received funding from any institution, such as NIH, please inform us of the grant number to go into the funding section. We use the institution names to tag NIH-funded articles so they are deposited at PMC.

These proofs are for checking purposes only. They are not in final publication format. Please do not distribute them in print or online. Do not publish this article, or any excerpts from it, anywhere else until the final version has been published with OUP. For further information, see https://academic.oup.com/pages/authoring/journals

Figure resolution may be reduced in PDF proofs and in the online PDF, to manage the size of the file. Full-resolution figures will be used for print publication.

Action: check your manuscript information

Please check that the information in the table is correct. We use this information in the online version of your article and for sharing with third party indexing sites, where applicable.

Full affiliations Each unique affiliation should be listed separately; affiliations must contain only the applicable department, institution, city, territory, and country.	NA
Group contributors The name of the group and individuals in this group should be given, if applicable (e.g. The BFG Working Group: Simon Mason, Jane Bloggs)	NA
Supplementary data files cited	NA
Funder Name(s) Please give the full name of the main funding body/agency. This should be the full name of the funding body without abbreviation or translation, if unsure, see https://search.crossref.org/funding	NA

Briefings in Bioinformatics, 2023, 1–13
https://doi.org/10.1093/bib/bbad390
Review

Identifying local associations in biological time series: algorithms, statistical significance, and applications

AQ5

Dongmei Ai, Lulu Chen, Jiemin Xie, Longwei Cheng, Fang Zhang, Yihui Luan, Yang Li, Shengwei Hou, Fengzhu Sun and Li Charlie Xia

Corresponding author. Department of Statistics and Financial Mathematics, School of Mathematics, South China University of Technology, Guangzhou 510641, China. E-mail: lcxia@scut.edu.cn

Abstract

Local associations refer to spatial–temporal correlations that emerge from the biological realm, such as time-dependent gene co-expression or seasonal interactions between microbes. One can reveal the intricate dynamics and inherent interactions of biological systems by examining the biological time series data for these associations. To accomplish this goal, local similarity analysis algorithms and statistical methods that facilitate the local alignment of time series and assess the significance of the resulting alignments have been developed. Although these algorithms were initially devised for gene expression analysis from microarrays, they have been adapted and accelerated for multi-omics next generation sequencing datasets, achieving high scientific impact. In this review, we present an overview of the historical developments and recent advances for local similarity analysis algorithms, their statistical properties, and real applications in analyzing biological time series data. The benchmark data and analysis scripts used in this review are freely available at http://github.com/labxscut/lsareview.

Keywords: Time series data; Local association; Local alignment; Local similarity analysis; Local trend analysis; Statistical significance

INTRODUCTION

Identification of ecological and biochemical relationships in time-series data of a microbiome or a transcriptome is commonly achieved by correlation-based co-occurrence and co-expression analyses. However, global correlation measures such as Pearson's or Spearman's correlation are only effective when two time series are globally synchronized and associated. Yet, more complex relationships with dynamic changes are prevalent in real biological data, including localized and time-delayed associations, which have been observed in various fields such as microbial ecology [1–4], molecular biology [5, 6], and functional neuroscience [7, 8].

For instance, the regulatory mechanism in gene expression analysis between ARG2 (acetyl glutamate synthase) and CAR2 (ornithine aminotransferase) depends on the expression level of CPA2 (arginine specific carbamoyl-phosphate synthase). Nonetheless, the global Pearson's correlation coefficient (PCC) expression level between ARG2 and CAR2 is nearly zero [9], which emphasizes the insensitivity of global correlation measures to detect local associations. Although global correlations continue to be valuable tool, it is necessary to formulate more flexible alternative hypotheses to detect the intricate relationships that exist in biological time series.

To address the limitations of global correlations, researchers have introduced numerous methodologies, the prominent one of such is local similarity analysis (LSA) [10–13]. LSA is a local alignment-based method that identifies alignment configurations optimized for local similarity scores between two time series, which enables the detection of local and potentially time-delayed pairwise associations. A related method, known as local trend analysis (LTA) [14–16], a variant of LSA, identifies alignment configurations and correlations between pairs of trend-of-change series, representing up, down, and no-change statuses. Both LSA

Dr. Dongmei Ai is a professor in the School of Mathematics and Physics at University of Science and Technology Beijing. Her research is focused on bioinformatics and computational biology.
Lulu Chen is a master's student in the School of Mathematics and Physics at University of Science and Technology Beijing. Her research interests include bioinformatics and computational biology.
Jiemin Xie is a master's student in the School of Mathematics, South China University of Technology. His research interest is bioinformatics.
Longwei Cheng is a master's student in the School of Mathematics and Physics at University of Science and Technology Beijing. His research interests include computational biology and machine learning.
Dr. Fang Zhang works as a machine learning specialist at the Shenwan Hongyuan Securities Co. Ltd. He received his PhD in Mathematics at the School of Mathematics, Shandong University. His research interests are statistics and machine learning.
Dr. Yihui Luan is a professor at the School of Mathematics at Shandong University. He works in the fields of biostatistics and time series analysis.
Yang Li is a master's student in the School of Mathematics, South China University of Technology. His research interest is deep learning and its application to biomedicine.
Dr. Shengwei Hou is an assistant professor in the Department of Ocean Science and Engineering, Southern University of Science and Technology. His research interests are bioinformatics, microbial ecology and evolution.
Dr. Fengzhu Sun is a professor of Quantitative and Computational Biology and Mathematics at the University of Southern California. He works in computational biology, bioinformatics, statistical genetics, and mathematical modeling.
Dr. Li Charlie Xia is a professor in the Department of Statistics and Financial Mathematics, South China University of Technology. He specialized in statistical modeling and algorithm development for biomedical data.

Received: May 25, 2023. **Revised:** August 21, 2023. **Accepted:** September 14, 2023

and LTA also evaluate the statistical significance of the identified associations. Notably, the term local similarity has not only been used in biology [17], but also in the fields of chemistry [18], public health [19], and power engineering [20], differing in meanings. We refer the readers to their respective papers and reviews.

This paper exclusively delves into the context of local similarity within biological time series. Within the realm of biological time series, local similarity analysis plays a pivotal role in uncovering potential relationships between microbial species richness and environmental factors. Local similarity becomes particularly relevant when actual correlations are limited to specific time subintervals or when correlations are subject to time delays [10]. The utilization of local similarity analysis has garnered extensive application in scrutinizing the spatiotemporal evolution of microbial communities across diverse environments. Importantly, the latent relationships between these microbial species cannot be effectively discerned through conventional correlation-based analysis methods.

Other methods exist for identifying diverse types of complex associations [9, 21–23]. For example, the liquid association (LA) technique [9, 22, 24] detects triplet rather than pairwise associations, while the maximal information coefficient (MIC) method identifies unspecified generic pairwise associations [25]. Granger causality analysis allows for identification of potential causal relationships in molecular ecological time-series data [26]. Furthermore, LA combined with LSA can be utilized to detect mediated covariant dynamics to discover the third-party mediator variable [27]. In this review, we focus on the advancement of LSA and LTA, including algorithms, statistical significance evaluation methods, software tools availability and their applications. Papers [28, 29] provide further information regarding related correlation measures and their applications.

LSA is founded on the concept of local sequence alignment, utilizing a dynamic programming algorithm, similar to the Smith-Waterman algorithm [17] to detect the optimal time series alignment configuration that maximizes the local similarity score. Qian et al. first introduced the method for gene expression profile analysis [10]. Based on this method, LSA has been extended to analyze the microbiome data including operational taxonomic unit (OTU) analysis by Ruan et al. [13] and metagenomic data analysis by Xia et al. [12]. Methodological advances have also been made to LSA, including support for replicate series data by Xia et al. [12], theoretical approximations of statistical significance (i.e., p-values) reported by Xia et al. [11] and Durno et al. [30], and more advanced methods for calculating p-values, such as moving block bootstrap LSA (MBBLSA) [31] and data-driven LSA (DDLSA) by Zhang et al. [32].

Accompanied by many other recently developed correlation measurement techniques for extracting meaningful association patterns from time series data, LSA by Ruan et al. [13] and the extended local similarity analysis tool, eLSA, a significantly accelerated implementation by Xia et al. [11, 12] remain advantageous. For instance, Wang et al. [33] developed a count statistic and evaluated its performance in comparison with LSA, the outcome revealed that as noise increased for pairs of correlated subsequences, LSA remained consistently robust and become more effective. Similarly, Tackmann et al. [34] introduced a Flashweave algorithm designed to infer high-resolution interaction networks while it was outperformed by eLSA when benchmarked with realistic data simulated by Weiss et al. [28].

Since 2011, the eLSA tool has been widely used in microbial data analyses, leading to numerous discoveries. For instance, Liu et al. [35] analyzed with eLSA the correlations between bacterial OTUs and phytoplankton in a reservoir and obtained a dynamical interaction network with significant local ecological associations. Similarly, Thiriet-Rupert et al. [36] constructed a gene co-expression network for a strain of microalgae to determine the relationships between differentially expressed transcription factors and mutant phenotypes by eLSA. Lee et al. [37] utilized eLSA to evaluate the time-dependent relationships among rhizome repair, environmental conditions, and bacterial genera in diesel-contaminated soil and obtained the correlation network.

Moreover, eLSA has been employed in studies of diverse subjects, including to investigate the correlation between marine paleontology and microbial communities by Parada et al. [38], to infer the diverse ecological relationships of microorganisms by Jones et al [39], to analyze the association network of microbial taxa of digested sludge by Liang et al. [40], and to identify the complex relationships among marine phytoplankton, bacteria and viruses [41, 42], between viral bacteriophages and their hosts [43] and between the abundance of antibiotic resistance genes and bacterial communities by many other researchers [44].

Additionally, eLSA's capability of identifying correlations with time shifts has also been widely used. For instance, Posch et al. [45] constructed an interaction network between ciliates and phytoplankton with eLSA to study the correlations with a time lag. They detected such associations among ciliates as well as between ciliates and algae. Džunková et al. [46] investigated the correlation among bacterial OTUs in the oral cavity of different individuals, considering the temporal delay. In another example, eLSA was utilized to evaluate the time-delayed correlation between microbial members of an activated sludge and environmental variables [47].

LSA related studies encouraged the development of methods sharing its rationale. One such method is local trend analysis (LTA), as introduced by He et al. [14] and Ji et al. [15], which analyzes the trend-of-change in gene expression profile series. In LTA, a local alignment algorithm is applied to the trend-transformed series of original data. Trend transformation involves taking the value differences of adjacent time points and discretizing them into {1, 0, −1} which represent {up, no-change, down} trending statuses, respectively. Xia et al. [16] developed a theory to approximate the statistical significance of LTA, which was further improved by Shan et al. known as the stationary theoretical local trend analysis (STLTA) [48].

Numerous studies have demonstrated the efficacy of LSA and LTA in inferring time-dependent associations in contrast to traditional global correlation measures. The added time dimension facilitates mining causality behind the delay. Additionally, the highly efficient theoretical approximation of statistical significance, i.e., p-value, has played a critical role in the widespread use of LSA and LTA analyses. These methods have been successfully applied to large-scale next-generation sequencing (NGS)-based datasets that would otherwise be impractical to analyze using much slower permutation-based p-value. In this paper, we provide an in-depth of the relevant algorithms and theories for LSA and LTA, as well as their recent improvements, including eLSA, fastLSA, MBBLSA, DDLSA, and STLTA.

LOCAL SIMILARITY ANALYSIS OF BIOLOGICAL SERIES DATA

Biological time series data

Time-series data are critical resources for studying the dynamics of biological systems. These data may be the gene expression levels in gene regulation studies or environmental and organismal factor levels in ecological studies. The nature of time-series data

Algorithm 1: Local Similarity Analysis (LSA)

```
• Input:
    – A pair of time series $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2} \ldots \mathbf{X_n})$ and $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2} \ldots \mathbf{Y_n})$
    – Delay limit $\boldsymbol{D}$
• Output:
    – Local Similarity Score $\boldsymbol{S}$
(1) Initialize scoring matrices $\boldsymbol{P}$ and $\boldsymbol{N}$:
    – For $\mathbf{i}, \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{n} : P_{0,i} = P_{j,0} = \mathbf{0}$ and $N_{0,i} = N_{j,0} = \mathbf{0}$.
(2) Carry out dynamic programming with restrictions:
    – For $\mathbf{i}, \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{n}$ with $|i - j| \leq D$:
      $P_{i+1,j+1} = \max\left\{0, P_{i,j} + X_{i+1} * Y_{j+1}\right\}$,
      $N_{i+1,j+1} = \max\left\{0, N_{i,j} - X_{i+1} * Y_{j+1}\right\}$.
(3) Compute the local similarity score corresponding to the
    alignment:
    – $P_{max} = \max_{1 \leq i,j \leq n} P_{i,j}$ and $N_{max} = \max_{1 \leq i,j \leq n} N_{i,j}$
    – $S = \frac{sgn(P_{max} - N_{max}) * \max(P_{max}, N_{max})}{\sqrt{n}}$
```

Note: *sgn(.)* is the sign function.

Algorithm 2: Extended Local Similarity Analysis (eLSA)

```
• Input:
    – A pair of replicated time series: $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2} \ldots \mathbf{X_n})$ and
      $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2} \ldots \mathbf{Y_n})$
    – Delay limit $\boldsymbol{D}$
• Output:
    – Local Similarity Score $\boldsymbol{S}$
(1) Initialize scoring matrices $\boldsymbol{P}$ and $\boldsymbol{N}$:
    – For $\mathbf{i}, \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{n} : P_{0,i} = P_{j,0} = \mathbf{0}$ and $N_{0,i} = N_{j,0} = \mathbf{0}$.
(2) Carry out dynamic programming with restrictions:
    – For $\mathbf{i}, \boldsymbol{j} = \mathbf{1}, \ldots, \boldsymbol{n}$ with $|i - j| \leq D$:
      $P_{i+1,j+1} = \max\left\{0, P_{i,j} + F(X_{i+1}) * F\left(Y_{j+1}\right)\right\}$,
      $N_{i+1,j+1} = \max\left\{0, N_{i,j} - F(X_{i+1}) * F\left(Y_{j+1}\right)\right\}$.
(3) Compute the local similarity score:
    – $P_{max} = \max_{1 \leq i,j \leq n} P_{i,j}$, and $N_{max} = \max_{1 \leq i,j \leq n} N_{i,j}$
    – $S = \frac{sgn(P_{max} - N_{max}) * \max(P_{max}, N_{max})}{\sqrt{n}}$
```

Note: *sgn(.)* is the sign function.

can be highly variable with values representing different measurements depending on the dataset.

Before conducting LSA on the original dataset, one needs to normalize the data. This step transforms raw values into standardized values with a mean of zero and a variance of one. However, one obstacle in standardizing count-based time series data is the high frequency of zeros. Excluding zero values from normalization procedures may be preferential since such values may be due to limited detection sensitivity or random dropouts other than true absence. Various normalization schemes are available, such as Z-score normalization and percentile normalization. Xia et al. used percentile normalization with sparsity adjustment, which generally provides the best and most robust result across datasets, even for datasets with non-linear associations (see paper [12] for the detailed procedure).

Local alignment and similarity score of biological series data

To explain the LSA process, we consider two factors with normalized time-series data at levels $X = (X_1, X_2 \ldots X_n)$ and $Y = (Y_1, Y_2 \ldots Y_n)$. Without loss of generality, the local similarity (LS) score is determined by calculating the maximized absolute value of summation $\sum_{k=0}^{l-1} X_{i+k} Y_{j+k}$, defined as $S = \max_{0 \leq i,j,l \leq n; |i-j| \leq D} \left|\sum_{k=0}^{l-1} X_{i+k} Y_{j+k}\right|$. Here, $i + k \in I = [i, i + l - 1]$ and $j + k \in J = [j, j + l - 1]$ which represent the aligned intervals of series X and Y. The dynamic programming algorithm based on the Smith-Waterman algorithm [17] can identify these intervals, as shown in Algorithm 1.

This local alignment algorithm has unique features compared to Smith-Waterman's algorithm. First, the maximum time delay, D, is introduced, which restricts the dynamic programming results to alignments with starting points in$X$ and Y that are maximally D time units apart. D is typically small since delayed associations usually represent immediate responses to recent changes in other factors, with a lag time that is generally short. A small D also effectively reduces the algorithm's search space, reducing the algorithm's computational and space complexities from quadratic to linear of the input size. Second, other than optimizing one objective, two score matrices, P and N, are simultaneously updated for maximization: one for positive and the other for negative associations. The local similarity score, S, is calculated as the maximum of all entries in P and N in absolute terms. S is then standardized by taking the square root of the series $\sqrt{n}$.

Local similarity score was initially scaled by n, as in refs. [11–13]. However, in retrospect, $\sqrt{n}$ is more desirable when comparing the scores across varied series lengths [26]. Henceforth, we will refer to this $\sqrt{n}$ scaled alignment score as the local similarity score or LS score. It should be noted that unlike Pearson's or Spearman's correlation, this LS score is not bound within the range of −1 to 1.

Local similarity score of replicated series data.

Assessing variability through replication is important for statistical inference [49–51]. The original LSA method only considered series data without replications. Xia et al. first proposed the extended local similarity analysis (eLSA) method to incorporate series data with replications [12]. The eLSA assumes that each sample has m replications, and a function $F\ (\cdot)$ is employed to summarize repeated measurements. The modified eLSA dynamic programming algorithm for repeated measures is presented in Algorithm 2.

In this modified algorithm, the scalars X_i and Y_i from the original LSA algorithm are replaced with summarized and transformed values $F(X_i)$ and $F(Y_j)$, respectively, which are applied to the m-replicated data vectors X_i and Y_j of size 1 by m. Therefore, the eLSA framework considers the original LSA as its special case when $m = 1$.

Transformation functions to summarize replicates

To model the repeated measure and account for variability in eLSA, Xia et al. introduced the summarizing function $F(\cdot)$. Researchers previously proposed several methods [51–53] to accomplish this task, including the simple mean method ('Mean'), the Standard Deviation-weighted mean method ('SD') [52], and AQ6 the multivariate correlation coefficient method [54], which is identical to the 'Mean' method. Moreover, compared to traditional methods, robust statistics do not need the normality assumption. They demonstrate effective resilience against the impact of outliers in the data, offering a more accurate reflection of the actual situation [54]. For instance, when employing the sample mean for overall mean estimation, the presence of an outlier

Table 1: Common summarizing functions $F(\cdot)$ for eLSA

Summarizing Method	$F(\cdot)$
Mean	$F(X_i) = \overline{X_i}$
SD	$F(X_i) = \frac{\overline{X_i}}{\sigma_{X_i}}$
Med	$F(X_i) = Median(X_i)$
MAD	$F(X_i) = \frac{Median(X_i)}{Median(\lvert X_i - Median(X_i)\rvert)}$

significantly influences the estimation outcome. On the contrary, the sample median exhibits lower sensitivity to data outliers, rendering it considerably more robust than the sample mean [55]. Therefore, robust statistics, such as Median and Median Absolute Deviation are robust replacement to the 'Mean' and 'SD' [56], and the resulting summary functions were included in eLSA and termed the median method ('Med') and the MAD-weighted median method ('MAD'), respectively. Table 1 shows the formula for common summarizing functions.

~~Bootstrap~~ confidence interval for local similarity score

The bootstrap method provides an empirical confidence interval (CI) for local similarity score S. For a given type-I error α, the estimated $1-\alpha$confidence interval will cover the true value of S with probability $1-\alpha$. The bootstrap samples $\overset{\sim}{X_i}^{k}$ and $\overset{\sim}{Y_j}^{k}$ are generated by sampling with replacement from the original data vectors X_i and Y_j, respectively, for a pre-specified total bootstrap sample number B. The eLSA calculation is performed on the bootstrap samples in the same way as it is performed on the original data, resulting in B bootstrap LS scores. These scores are then sorted in ascending order, such that $\tilde{S}^{(k)} = S\left(\tilde{X}^{(k)}, \tilde{Y}^{(k)}\right)$ represents the k-th ranked score. The 95% CI for the original score S is inferred as $\left[\tilde{S}^{(\lfloor \frac{\alpha}{2}B \rfloor)}, \tilde{S}^{(\lfloor 1-\frac{\alpha}{2}B \rfloor)}\right]$.

A statistical theory for the significance of local similarity score

Permutation statistical significance

Non-replicated data

In the original non-replicated data LSA analysis, the statistical significance (p-value) of an LS score was computed by permutation. This process involved the following steps:

(i) For each pair of time series X and Y with LS score $S(X,Y)$, permute the data values in each sequence to generate a permuted sample pair.

(ii) Compute a permuted LS score for the permuted pair with the LSA algorithm (Algorithm 1).

(iii) Repeat steps (i) and (ii) L times to obtain $(X^1, X^2, \ldots, X^L)$ and $(Y^1, Y^2 \ldots Y^L)$, which represent L permuted samples of X, Y and compute their permuted LS scores.

(iv) The p-value is determined as the proportion of permuted LS scores that were at least as great as $S(X,Y)$, i.e.,

$$P_L = Prob\left[S \geq S(X,Y)\right] \approx \frac{1}{L}\sum_{k=1}^{L} I\left[S\left(X^k, Y^k\right) \geq S\left(X,Y\right)\right], \tag{1}$$

where $I(\cdot)$ is the indicator function.

Replicated data

In the case of replicated data eLSA analysis, the permutation test is modified such that the series data Y remains unchanged for each permuted sample while all columns of X are reshuffled. For a fixed number of permutations L, and $(X^1, X^2 \ldots X^L)$ are the permuted samples of X, and the p-value P_L is computed by:

$$P_L = Prob\left[S \geq S(X,Y)\right] \approx \frac{1}{L}\sum_{k=1}^{L} I\left[S\left(X^k, Y\right) \geq S\left(X,Y\right)\right]. \tag{2}$$

Moving block bootstrap

Zhang et al. [31] developed a moving block bootstrap method (MBBLSA) to replace point-wise permutation and compute the statistical p-value of LSA. The MBBLSA calculation process is as follows:

(i) Calculate the LS scores $S\ (X,Y)$ for each pair of time series X and Y.

(ii) Divide the time series into $n-l+1$ continuous time block of equal length, assuming the time series length is n and the block length is l for each block.

(iii) Randomly select a block from all of the blocks and use its values as the first l resamples. Then, continue selecting blocks randomly, using the next l values as the next resample. If the k-th permuted sample $X^k = (X_1^k, X_2^k, \ldots X_n^k)$ has already generated data points at n time points, discard the data points after X_n^k and proceed to step (iv). Otherwise, continue sampling.

(iv) Calculate the LS scores of X^k and Y, and record them as $S(X^k, Y)$.

(v) Repeat steps (iii) and (iv) to obtain $(X^1, X^2, \ldots, X^L)$ and its LS score $(S(X^1, Y), S(X^2, Y), \ldots, S(X^L, Y))$.

(vi) Calculate the bootstrap p-value as in Eq. (2).

Regarding the block length, please refer to the block length selector proposed by Sherman et al. [57]:

$$l = NI\left\{\left[6^{\frac{1}{2}} \bullet \frac{\hat{a}}{1-\hat{a}^2}\right]^{\frac{2}{3}} \bullet n^{\frac{1}{3}}\right\}, \tag{3}$$

where $NI(\bullet)$ is the nearest integer function, and $\hat{a}$ is the estimated value of the autoregressive coefficient obtained by fitting the autoregressive (AR) (1) model with data.

Bottlenecks of permutation and bootstrap procedures

Although the p-value can be computed with any desired degree of accuracy using a large number of permutations, this approach can be computationally costly. The time complexity for calculating the LS score is $\Theta(n)$ at runtime for a single pair of series, where n is the number of time points. The time complexity for computing bootstrap CI via B bootstraps for the LS score is $\Theta(Bn)$. The time complexity for estimating the statistical significance of a pair of series through L permutations is $\Theta(Ln)$. Thus, the time complexity for conducting a full eLSA analysis on a single pair of series is$\Theta(BLn)$. With series data of T factors, the total pairs are $\frac{T(T-1)}{2}$. Then for an all-to-all analysis, the total time complexity will be $\Theta(T^2BLn)$, which is impractical for large-scale datasets with thousands of factors, such as genome-wide multi-omics or metagenomics microbiome datasets. This motivates the demand for the development of theory-based approaches for efficient p-value approximation.

Theoretical statistical significance for local similarity scores

Feller's theory for the range of random walk excursion

Due to computational complexity, permutation-based procedures are only practical for pairwise LSA analysis with a few hundred factors. Therefore, it is necessary to develop effective theoretical p-value approximations to alleviate this problem. One direction involves adapting Feller's theory for the range of random walk excursions to the settings of LSA. Feller [58] investigated the approximate distribution of the range of the sum of $\mathbf{n}$ random variables with an expectation of zero. Let Z_i be independent and identically distributed (*i.i.d.*) random variables such that $E(Z_i) = 0$ and $Var(Z_i) = \sigma^2$, where $E(\bullet)$ is the mathematical expectation, and $Var(\bullet)$ is the variance. Let $C_n = Z_1 + Z_2 + \cdots + Z_n$, $M_n = \mathbf{max}\{0, C_1, C_2, \cdots, C_n\}$, and $m_n = \mathbf{min}\{0, C_1, C_2, \cdots, C_n\}$. The excursion range is then defined as $R_n = M_n - m_n$. According to Feller [56], it can be concluded that:

$$P(S_n \geq x) = P\left(\frac{R_n}{\sigma\sqrt{n}} \geq x\right) = 1 - 8\sum_{k=1}^{\infty}\left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right) \mathbf{exp}\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \quad (4)$$

A theory for local similarity score

Xia et al. [11] proposed the LS score, S_n, which is computed through the dynamic alignment of a pair of standard normalized series with length n. It was established that the LS score follows exactly the distribution of $\frac{R_n}{\sigma\sqrt{n}}$ when $\sigma = 1$. Notably, Eq. (4) involves an infinite sum, and a stopping rule is required for numerical approximation. In Xia et al. [11], an upper bound was established to determine when to terminate the summation for practical p-value calculations. For a specified error limit, β, one can choose a summation term upper bound, K, such that:

$$\frac{16\,\mathbf{exp}\left(-\frac{(2K+1)\pi^2}{2x^2}\right)}{x^2\left(1 - \mathbf{exp}\left(-\frac{2\pi^2}{2x^2}\right)\right)} \leq \beta. \quad (5)$$

Then, the tail probability $P\left(\frac{R_n}{\sigma\sqrt{n}} \geq \mathbf{x}\right)$, i.e., the p-value of S_n, is approximated by:

$$P(S_n \geq x) = P\left(\frac{R_n}{\sigma\sqrt{n}} \geq \mathbf{x}\right) \approx 1 - 8\sum_{k=1}^{K-1}\left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right) \mathbf{exp}\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \quad (6)$$

A generalization of the theory of local similarity scoring

The approach proposed by Xia et al. [11] is effective only for independent and identically distributed (*i.i.d.*) sequences. However, in practice, many sequences are not *i.i.d.*. To solve this issue, Zhang et al. [32] introduced the long-term variance ω^2 and replaced σ with ω to approximate the p-value using Eq. (4).

Accurately estimating ω is crucial for evaluating the statistical significance of LS scores for dependent sequences. Zhang et al. [32] applied the data-dependent $AR(1)$ plug-in [59] method to estimate long-term variance as:

$$\omega^2 = \hat{\gamma}_x(0)\hat{\gamma}_y(0) + 2\sum_{k=1}^{b_\omega}\left(1 - \frac{k}{b_\omega}\right)\hat{\gamma}_x(k)\hat{\gamma}_y(k), \quad (7)$$

where $\hat{\gamma}_x(k)$ and $\hat{\gamma}_y(k)$ are the sample auto-covariance functions of X and Y, respectively. The bandwidth parameter, b_ω, should be solved from $b_\omega = 1.447(\hat{\tau}n)^{\frac{1}{3}}$ [57], where:

$$\hat{\tau} = \frac{4\hat{\phi}^2}{\left(1 - \hat{\phi}^2\right)^2}, \quad \hat{\phi} = \frac{\sum_{i=2}^{n}\hat{u}_t\hat{u}_{t-1}}{\sum_{i=2}^{n}\hat{u}_t^2}, \quad \hat{u}_t = Z_t - \overline{Z},$$

$Z_t = X_tY_t$, $t = 1, \ldots, n$, and $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is the mean of Z_t.

~~An upper bound approach for statistical significance~~

Durno et al. [30] proposed an alternative upper bound for the LS score p-value with the fastLSA software. They slightly modified the presentation of the LSA algorithm (Algorithm 1) to enable derivation of their bound for the tail distribution of the LS score statistics. They defined $Z_k = X_{i+k} * Y_{j+k}$, $k = 0, \ldots, \mathbf{min}\{n - i, n - j\} - 1$. They established the following equivalence to derive the tail probability bounds: $\{|S| > x\} = \left\{\left(U_{i,j}\{P_{i,j} > nx\}\right) \cup \left(U_{i,j}\{N_{i,j} > nx\}\right)\right\}$ (Lemma 1 in ref [30], where U represents the union operation). This equivalence can be used to express the probability of the tail event $\{|S| > x\}$ (i.e., the targeted p-value) in terms of the probabilities of the tail events $\{P_{i,j} > x\big\}$ and $\{N_{i,j} > x\}$.

For aligned sequences starting from X_i and Y_j, Durno et al. found that $\{Z_k\}$ are *i.i.d.* random variables. Consequently, condition established by Lindeberg [15] was fulfilled for $\{Z_k\}$ (Lemma 2 in ref. [30]), which means that the variance of a distribution stabilizes as more variables are incorporated, and its tails are pinned down. Essentially, as the time series get longer, the upper bound of the distribution becomes more well-defined and calculable. In other words, if the random variables $\{Z_k\}$ have zero expectation and finite variance and the Lindeberg's condition holds, its distribution uniformly converges to that of a one-sided standard normal distribution as the length of the sequence becomes infinite (Theorem 3 in ref. [30]). By applying this theorem to the $\{Z_k\}$ sequence, they obtained the following:

$$P(|S| > x) \leq 2\left(n^2 - (n - D - 1)(n - D)\right)\left(1 - G\left(x\sqrt{\frac{n}{Var(X_1Y_1)}}\right)\right), \quad (8)$$

where $G(x) = \sqrt{\frac{2}{\pi}}\int_0^x e^{-t^2/2}dt$ if $x \geq 0$and 0 if $x < 0$, which represents the cumulative distribution function of one-sided standard normal distribution (Theorem 4 in ref. [30]).

Durno et al. acknowledged the limitations of their approach. Note that Eq. (8) is asymptotic, and n must be substantially large for this bound to be applicable. They suggested that at least 30 time points are needed for reliable LS score p-value bound implementation. Additionally, it should be noted that convergence can vary across datasets, and applying fastLSA to shorter time series can produce non-negligible false positives, despite the bound being conservative.

A comparison of statistical significance approximation approaches

We compared five available open-source LSA software tools that implement either the permutation or theoretical approaches for estimating the statistical significance of LS scores, as presented in Table 2. The tools include eLSA [11, 12], fastLSA [30], LocSim [13], MBBLSA [31] and DDLSA [32]. Both eLSA and fastLSA are implemented in efficient C++ code, while LocSim, MBBLSA, and DDLSA are written in R. eLSA offer both theoretical (Xia et al. [11]) and permutation-based p-value calculations. FastLSA and DDLSA provide the theoretical p-value calculation. Meanwhile, LocSim and MBBLSA only implement the permutation approach. eLSA also includes additional features, such as trend series analysis, replicated data analysis, and bootstrap confidence intervals. Furthermore, we provided Figure 1 as a viable decision tree for selecting appropriate software tools for various tasks or datasets.

Furthermore, we compared the performance of p-value approximation approaches: theory or permutation. The theoretical methods were implemented in eLSA (Xia et al. [11]), fastLSA

Table 2: Feature comparison of local similarity analysis software tools

	LocSim	eLSA	fastLSA	MBBLSA	DDLSA
Reference	Ruan et al. 2006 [13]	Xia et al. 2011,2013 [11, 12]	Durno et al. 2013 [30]	Zhang et al 2018 [31]	Zhang et al 2019 [32]
Homepage	https://www-rcf.usc.edu/&#x007E;fsun/Programs/local_similarity/lsaMain.html	https://github.com/labxscut/elsa	http://www.cmde.science.ubc.ca/hallam/fastLSA	https://github.com/BlueStamford/MBBLSA	https://github.com/BlueStamford/DDLSA
Language	R	Python and C++	C++	R	R
Replicated Data	No	Yes	No	Yes	Yes
Confidence Interval	No	Yes	No	No	No
p-value	Permutation	Theory and Permutation	Theory	Permutation	Theory

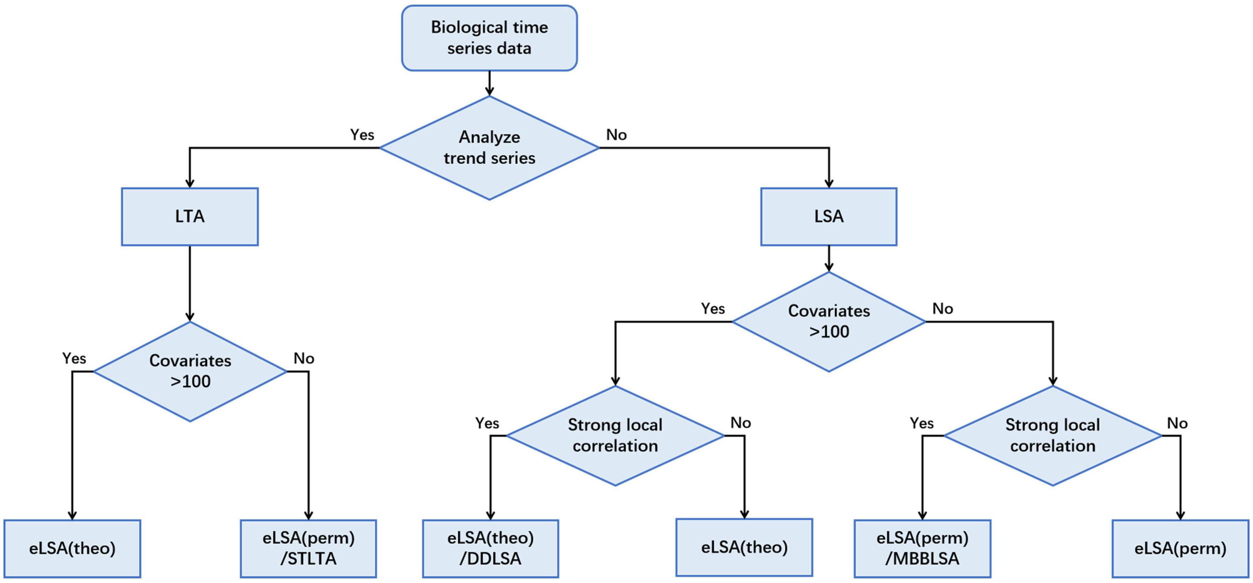

Figure 1. Decision tree for choosing LSA software tools.

(Durno et al. [30]) and DDLSA (Zhang et al. [32]). The permutation approaches were implemented in LocSim (Ruan et al. [13]) and MBBLSA (Zhang et al. [31). The LocSim R tool was excluded from this comparison because its permutation approach is the same as that in eLSA but slower.

We obtained the benchmark data prepared by Weiss et al. from the *GitHub* (https://github.com/wdwvt1/correlations) [28]. For the benchmark purpose of this paper, we utilized a portion of their data which is shown in Table 3. Tables S1-4 (originally Tables 2.6–2.9 in Weiss et al.) simulated the *Lotka-Volterra* relationship of two species with different constraints. The *Lotka-Volterra* relationships describe *n* species' abundances by a system of *n* differential equations that mimic ecological relationships. The simulated positive *Lotka-Volterra* relationships in these tables were confounded with simulated random covariates (i.e., OTUs) from the lognormal and gamma distributions as negative controls.

After pre-processing the raw tables, we filtered out sequences with more than 80% time points being zero values. In the final analysis, Tables S1 and S3 contained 4770 pairs of sequences, including 10 pairs of time series with true correlation. Table S2 contained 4366 pairs of sequences, including 6 pairs of time series with true correlation, and Table S4 contained 4285 pairs of total sequences, including 5 pairs of time series with true correlation. We then input these series data into each LSA analytical software tool to calculate the pairwise LS scores between covariates.

We conducted benchmark tests on the two-species Lotka-Volterra (TLV) datasets with the eLSA, fastLSA, MBBLSA, and DDLSA tools. The performance of each tool was illustrated in Figure 2 by the Receiver Operating Characteristic (ROC) curve (the *pROC* R package). The curves showed that the eLSA theory (theo), eLSA permutation (perm), and DDLSA methods generally outperformed others in the tests. Specifically, the eLSA (theo) method demonstrated the highest overall accuracy compared to all other methods. Furthermore, we evaluated the tools with data simulating many-species *Lotka-Volterra* relationship and compared the resultant ROC curves, as shown in Supplementary Table S1 and Figure S1. We observed that the DDLSA methods yielded the best performance in those tests.

Furthermore, Table 4 shows a ranking of correctly identified sequence pairs by all the tested software tools. In the table, *n* represents the total number of OTU pairs calculated by the tool, and the *accurate* column recorded the number of correctly called pairs by the tool. Among these tools, we observed that the eLSA (theo) method exhibited the best performance and ranked the first in overall accuracy.

Table 3: Summary of simulation models for the benchmark data

Data File	Data Name	Simulating Model	Data Source
Tables S1	TLV-even-relative	Two-species Lotka-Volterra(TLV), even indices, relative abundance	Weiss 2016 [28] Table Set 2.6
Tables S2	TLV-even-counts	Two-species Lotka-Volterra(TLV), even indices, counts	Weiss 2016 [28] Table Set 2.7
Tables S3	TLV-random-relative	Two-species Lotka-Volterra(TLV), random indices, relative abundance	Weiss 2016 [28] Table Set 2.8
Tables S4	TLV-random-counts	Two-species Lotka-Volterra(TLV), random indices, counts	Weiss 2016 [28] Table Set 2.9

Note: Tables S1-S10 can be obtained at http://github.com/labxscut/lsareview.

Table 4: Total pairs correctly identified by the LSA software tools in benchmark (p-value <0.05)

	TLV-even-relative (n = 4770)		TLV-even-counts (n = 4366)		TLV-random-relative (n = 4770)		TLV-random-counts (n = 4285)	
	#accurate	rank	#accurate	rank	#accurate	rank	#accurate	rank
eLSA(perm)	4486	5	4148	4	4466	4	4055	5
eLSA(theo)	4745	1	4356	1	4731	1	4232	1
fastLSA	4615	3	4201	3	4660	2	4145	3
DDLSA	4696	2	4285	2	4654	3	4183	2
MBBLSA	4507	4	4131	5	4465	5	4056	4

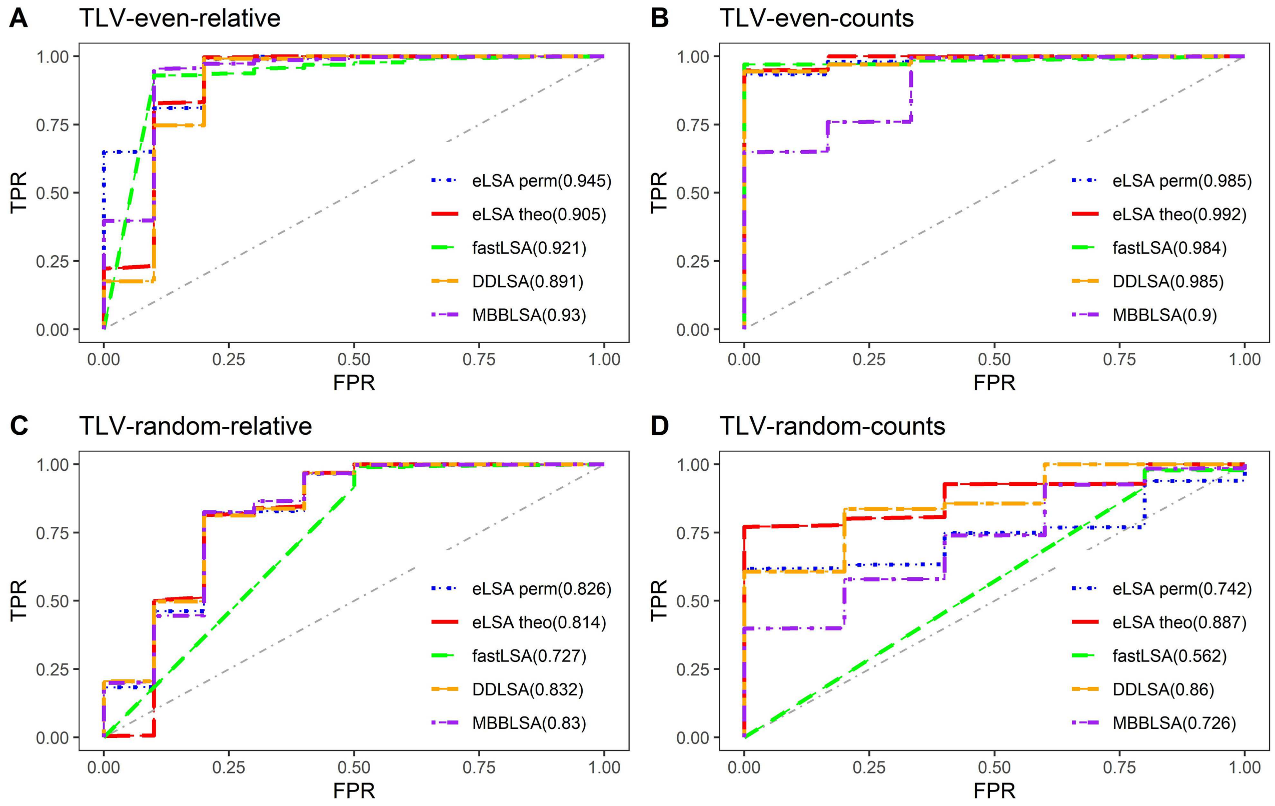

Figure 2. ROC curves comparing four LSA software tools and their implemented P-value approximation approaches. Panels **A**–**D** show the area under the curve (AUC) scores of eLSA, fastLSA, DDLSA and MBBLSA under different data simulation models. The blue curve represents eLSA with permutation (eLSA-perm), and the red curve represents eLSA with theory approximation (eLSA-theo). Meanwhile, fastLSA (green) and DDLSA (orange) are theoretical, and MBBLSA (purple) is permutation-based approximation. Further details about these approaches can be found in the main text.

AQ7

We also conducted a comparison of computational efficiency for the four LSA software tools in terms of the total running time, as measured in minutes, from the benchmark tests (see Table 5). The results indicated that all theoretical approaches - eLSA (theo), fastLSA and DDLSA, demonstrated good performance in terms of computational efficiency, showing more than hundreds of times acceleration than permutation. They finished the analysis within seconds when handling hundreds of covariates (OTUs). Our experience showed, the theory-based approaches can scale up to thousands of covariates using current single processor personal computer, while the permutation approaches become too slow to finish. Overall, the benchmark tests demonstrated that eLSA (theo) is the most recommended choice for LSA analysis, considering overall accuracy and efficiency.

Table 5: Running time (in minute) of the LSA software tools in benchmark

	TLV-even-relative (n = 4950)	**TLV-even-counts (n = 4465)**	**TLV-random-relative (n = 4950)**	**TLV-random-counts (n = 4371)**
eLSA(perm)	513.8	468.7	613.7	466
eLSA(theo)	1.8	1.3	1.7	1.3
fastLSA	<0.1	<0.1	<0.1	<0.1
DDLSA	0.2	0.2	0.2	0.2
MBBLSA	439.7	302	410	296.6

Local trend analysis and statistical significance

Local trend series and local trend analysis

A related LSA analysis technique, local trend analysis (LTA) was also implemented in eLSA for time-dependent trend association mining. Recent studies have suggested that the similarity of increasing, stabilizing or decreasing trends along the timeline can be a strong indicator of associations. To address this, trend series analysis was thus developed to analyse the transformed series. Ji et al. [15] explored this approach by coding the trends of n consecutive time points of gene expression profiles into an $n-1$ time point series representing the local trends of *increase*, *no change*, and *decrease* in gene expression levels. They analysed the transformed series using an exhaustive search algorithm to identify possible local associations of specific length or time span between genes. Later, He et al. [14] updated the analysis with a dynamic programming algorithm and a permutation p-value approach, identical to that of LSA, to compute similarity scores and evaluate the statistical significance. Therefore, we refer to the LSA technique when applied to the transformed trend series as local trend analysis or LTA, with its corresponding similarity measure as the local trend (LT) score.

In LTA, the initial step is to discretize the series profile using the changing trend alphabet $\boldsymbol{\Sigma}$, which reflects a set of symbols of interest representing distinct direction changes in states. Typically, a two-letter alphabet such as $\boldsymbol{\Sigma} = \{-1, 1\}$ for trend-down and trend-up states, or a three-letter alphabet such as $\boldsymbol{\Sigma} = \{-1, 0, 1\}$ for trend-down, no-change and trend-up states, is applied. It is also achievable to discretize into a bigger-size alphabet.

For the given original time series $X = (X_1, X_2 \ldots X_n)$, we transform the 1-by-n vectorX to a 1-by-$(n-1)$ trend vector d_i^X, $i = 1, 2, \cdots, n-1$ by the following rules: When$X_i \neq 0$,

$$d_i^X = \begin{cases} 1, & \mathit{if} \ \frac{X_{i+1}-X_i}{|X_i|} \geq t \\ 0, & \mathit{if} -t < \frac{X_{i+1}-X_i}{|X_i|} < t \\ -1, & \mathit{if} \ \frac{X_{i+1}-X_i}{|X_i|} < -t \end{cases}, \tag{9}$$

where $t \geq 0$ is a threshold value for declaring changing trends. When$X_i = 0$, $d_i^X$is simplified as

$$d_i^X = \begin{cases} 1, \ \mathit{if} \ \ X_i = 0 \text{ and } X_{i+1} > 0 \\ 0, \ \mathit{if} \ \ X_i = 0 \text{ and } X_{i+1} = 0 \\ -1, \ \mathit{if} \ \ X_i = 0 \text{ and } X_{i+1} < 0 \end{cases}. \tag{10}$$

Statistical significance of local trend score

Similar to LSA, a theory for the statistical significance of the LT score in LTA is necessary. After transformation, the resulting trend series is no longer independent and identically distributed (*i.i.d.*). As illustrated in Figure 3, it has a dependence structure wherein the dependence relationship is diminishing over time. Xia et al. [16] demonstrated that a first-order Markov chain provides a good approximation of the resulting dependent trend series in practice:

$$P\left(\left(d^X d^Y\right)_i | \left(d^X d^Y\right)_{i-1}, \ldots, \left(d^X d^Y\right)_1\right) \approx P\left(\left(d^X d^Y\right)_i \mid \left(d^X d^Y\right)_{i-1}\right). \tag{11}$$

Therefore, the LT score S, as determined by LTA, is identical to the LS score, except that it represents the maximum range of random walk excursion of the sum of Markov-like random variables.

The theory by Xia et al. for LS score can be extended to Markov random variables to approximate the p-value of LT score. Daudin et al. [87] explored the distribution of the maximum cumulative sum of first-order ~~Markov~~ chain with values taken from a finite subset of R. Let φ be the stationary distribution of such Markov random variables $Z_i, i = 1, 2, \cdots$, where $E\varphi \ (Z_1) = 0$ and

$$\sigma_\varphi^2 = E_\varphi\left(Z_1^2\right) + 2\sum_{k=1}^{\infty} E_\varphi\left(Z_1 Z_{k+1}\right). \tag{12}$$

Therefore, following Eq. (6), the p-value approximation formula for LT score is:

$$P\left(\frac{R_n}{\sigma_\varphi\left(d^X d^Y\right)\sqrt{n}} \geq x\right) \approx 1 - 8\sum_{k=0}^{K-1}\left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right) \exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \tag{13}$$

Eq. (13) is equivalent to Eq. (6), where the only difference being that the standard deviation σ_φ is no longer 1. σ_φ must be calculated from the stationary distribution φ of trend series $d^X d^Y$. In the case of a two-letter alphabet case (i.e., $t = 0$ and $\boldsymbol{\Sigma} = \{-1, 1\}$), an exact solution was obtained for σ_φ by Xia et al. [16], which is $\sqrt{1.25}$. However, in the case of three-letter alphabet (i.e., $t > 0$ and $\boldsymbol{\Sigma} = \{-1, 0, 1\}$), Monte Carlo simulations are necessary to approximate the stationary distributionφ numerically, and the spectral expansion technique is required to solve for σ_φ. Xia et al. [16] provides the details for such calculations, as well as the necessary numerical equations to handle multi-letter alphabets.

The analysis conducted by Xia et al. [16] assumed that the original sequence is *i.i.d.*. Shan et al. [48] mitigated this assumption by proposing the spectral decomposition theory of matrices, which enables the solution of σ_ϕ^2 to be expanded to a broader range of scenarios. When $t = 0$, it is assumed that d_i^X and d_i^Y are two independent two-state Markov chains, but the transition probability matrices of d_i^X and d_i^Y are assumed to be distinct 2-by-2 matrices T_X and T_Y:

$$T_X = \left[\begin{array}{ccc} state & -1 & 1 \\ -1 & a_X & 1-a_X \\ 1 & 1-a_X & a_X \end{array}\right], T_Y = \left[\begin{array}{ccc} state & -1 & 1 \\ -1 & a_Y & 1-a_Y \\ 1 & 1-a_Y & a_Y \end{array}\right], \tag{14}$$

where a_X and a_Y are the transition probabilities.

Under the condition that X_i and Y_i are independent, then:

$$\sigma_\varphi^2 = \frac{1+(2a_X-1)(2a_Y-1)}{1-(2a_X-1)(2a_Y-1)}, \tag{15}$$

was proposed by Shan et al. [48] as plug-in estimates to Eq. (13) for computing the LTA p-value. When $t \neq 0, d_i^X$ and d_i^Y are three-state Markov chain, the 3-by-3 transition probability matrices of d_i^X and d_i^Y are:

$$T_X = \left[\begin{array}{cccc} state & -1 & 0 & 1 \\ -1 & b_X & 1-b_X-c_X & c_X \\ 0 & d_X & 1-2d_X & d_X \\ 1 & c_X & 1-b_Y-c_X & b_X \end{array}\right],$$

$$T_Y = \left[\begin{array}{cccc} state & -1 & 0 & 1 \\ -1 & b_Y & 1-b_Y-c_Y & c_Y \\ 0 & d_Y & 1-2d_Y & d_Y \\ 1 & c_Y & 1-b_Y-c_Y & b_Y \end{array}\right]. \tag{16}$$

where b_i, d_i and c_i $\left(i = X, Y\right)$ are the transition probabilities.

If X_i and Y_i are independent,

$$\sigma_\varphi^2 = 4\left(\frac{d_X}{1-b_X-c_X+2d_X}\right)\left(\frac{d_Y}{1-b_Y-c_Y+2d_Y}\right) \left(\frac{1+(b_X-c_X)(b_Y-c_Y)}{1-(b_X-c_X)(b_Y-c_Y)}\right), \tag{17}$$

was proposed by Shan et al. [48] as plug-in estimates to Eq. (13) for computing the LTA p-value.

Shan et al. [48] also studied the mixed-state Markov chain model. When d_i^X is a two-state Markov chain and d_i^Y is a three-state Markov chain (referred to as a mixed-state Markov model), then:

$$\sigma_\varphi^2 = \left(\frac{2d_Y}{1-b_Y-c_Y+2d_Y}\right)\left(\frac{1+(2a_X-1)(b_Y-c_Y)}{1-(2a_X-1)(b_Y-c_Y)}\right), \tag{18}$$

the proposed plug-in estimates to Eq. (13) for computing the LTA p-value.

Shan et al. [48] referred to this method as STLTA. Through simulation, they demonstrated that the Type I error rate of the STLTA method closely aligns with the specified significance level as the number of time points increases, which indicates that the STLTA method is effective asymptotically.

Finally, Table 6 presented a summary of LSA and LTA software tools and their real applications in various data domains and specialties, aiming to provide practitioners with a convenient overview of the wide adoption of these tools.

DISCUSSION AND CONCLUSIONS

Local association analysis methods, such as local similarity and trend analysis (LSA and LTA), are powerful tools for identifying time-dependent associations in biological time series data. These methods are widely applied in hypothesis generation to determine the most relevant interactions in biological systems. Standard dynamic programming algorithms are often employed to find the optimal time series alignment and compute local similarity and trend scores. Advances in theoretical statistical significance approximation have addressed the common computational bottleneck faced by researchers, notably when analysing large-scale NGS datasets, LSA and LTA methods were brought under a common conceptual framework of random walk excursions for theoretical study. Nevertheless, computational and theoretical challenges remain in cases where local association has gaps, all-to-all comparing tens of thousands or even more variables, or analysing local association involves three or more co-factors. Further development of novel algorithms and statistical theories will be necessary enable these analyses in the future.

Currently, multiple software tools are accessible for executing LSA. Table 2 provides a clear comparison of the features of five such tools: eLSA, fastLSA, LocSim, MBBLSA, and DDLSA. Although the older R tool LocSim is limited to the permutation procedure for *p*-values, the recent Python and C++ implementation, i.e., eLSA, supports both local similarity and trend analysis and provides both permutation and theoretical approximation for *p*-values. eLSA is particularly suitable for large NGS-based datasets.

Table 6: Example real-world data analysis applying LSA and LTA software tools

Method (Software)	Domain	Data Source	Example Applications
LSA	Health Science	16S rRNA Seq	Refs. [46, 60–63]
(eLSA)		Metagenomics	Ref. [64]
	Food / Agriculture	16S rRNA Seq	Refs. [4, 44, 65–67]
	Ecology	Metagenomics	Refs. [43, 47, 68]
		RNA-Seq	Ref. [36]
		16S rRNA Seq	Refs. [41, 69–79]
	Bioenergy	16S rRNA Seq	Refs. [80, 81]
LSA	Ecology	16S rRNA Seq	Refs. [82–86]
(fastLSA)			
LTA	Molecular systems biology	RNA-Seq	Refs. [88, 89]
	Biomedicine	RNA-Seq	Ref. [90]

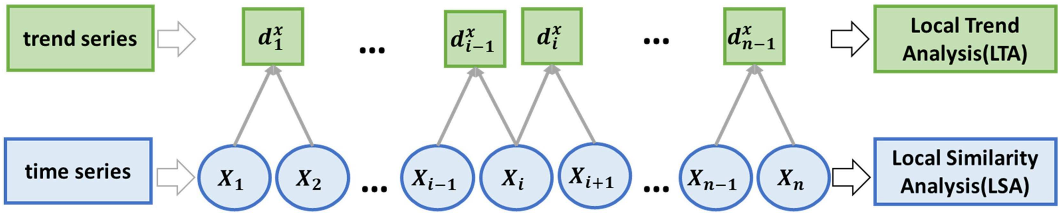

Figure 3. A generative illustration for local trend time series. The adjacent continuous values of the original time series are compared and thresholded to create a new trend series, which is a categorical series suitable for local trend analysis. This enables the identification of aligned directions of changes.

eLSA also supports replicated series with multiple summarizing functions and bootstrap confidence intervals. Recently developed MBBLSA and DDLSA methods further improved the permutation and theoretical approximation of p-values. Additionally, another C++ implementation, fastLSA, trades off p-value accuracy for faster computing speed.

This benchmark study has shown that eLSA is the preferred software tool for the analysis of time series data in both sparse and dense data scenarios. Weiss et al. [28] also recommended eLSA, along with MIC [25], as the best analysis methods for dense cross-sectional data when the delay D is set to zero. They highlighted the desirable robust properties of LSA (as implemented in eLSA), irrespective of the distribution and sparsity of the dataset. We recommend that readers refer to ref. [28] for a more in-depth understanding of the comprehensive evaluation therein. Moreover, ref. [29] provides a summary of various analytical methods for studying microbial community networks.

ABBREVIATIONS

DDLSA: Data-driven local similarity analysis
eLSA: Extended local similarity analysis.
i.i.d.: Independent and identically distributed.
LA: Liquid association.
LSA: Local similarity analysis.
LTA: Local trend analysis.
MBBLSA: Moving block bootstrap local similarity analysis.
MIC: Maximal information coefficient.
NGS: Next-generation sequencing.
OTU: Operational taxonomic unit.
PCC: Pearson's correlation coefficient.
STLTA: Stationary theoretical local trend analysis.

Key Points

- This paper provides a comprehensive review and analysis of local similarity analysis (LSA), including algorithms, statistical theories, and software tools.
- Various software tools' accuracy and efficiency are compared to aid readers in choosing the most appropriate tool.
- Alongside LSA, this paper also reviews local trend analysis (LTA) and its statistical significance theories.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENTS

We thank Drs. Jed Fuhrman, Jacob Cram and Joshua Steele for helpful discussions. We thank Drs. R. Knight, S. Weiss, and Mr. Van Treuren for developing and sharing the benchmark data tables.

COMPETING INTERESTS

The authors declare no competing interests.

FUNDINGS

This work was supported by grants from the National Natural Science Foundation of China (61,873,027 to D.A. and L.C.X.), Open Project of the National Engineering Laboratory for Agri-product Quality Traceability (No.AQT-2020-YB6 to D.A.), and Guangdong Basic and Applied Basic Research Foundation (2022A1515–011426 to L.C.X.). AQ8

REFERENCES

1. Caporaso JG, Lauber CL, Costello EK, *et al.* Moving pictures of the human microbiome. *Genome Biol* 2011;**12**:R50.
2. Cram JA, Xia LC, Needham DM, *et al.* Cross-depth analysis of marine bacterial networks suggests downward propagation of temporal changes. *ISME J* 2015;**9**:2573–86.
3. Steele JA, Countway PD, Xia L, *et al.* Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* 2011;**5**:1414–25.
4. Shade A, McManus PS, Handelsman J. Unexpected diversity during community succession in the apple flower microbiome. *MBio* 2013;**4**:e00602–12.
5. Cho RJ, Campbell MJ, Winzeler EA, *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998;**2**: 65–73.
6. Spellman PT, Sherlock G, Zhang MQ, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998;**9**:3273–97.
7. Amar D, Yekutieli D, Maron-Katz A, *et al.* A hierarchical Bayesian model for flexible module discovery in three-way time-series data. *Bioinformatics* 2015;**31**:i17–26.
8. Vaisvaser S, Lin T, Admon R, *et al.* Neural traces of stress: cortisol related sustained enhancement of amygdala-hippocampal functional connectivity. *Front Hum Neurosci* 2013;**7**:313.
9. Li KC. Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A* 2002;**99**:16875–80.
10. Qian J, Dolled-Filhart M, Lin J, *et al.* Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new. Biologically relevant interactions. *J Mol Biol* 2001;**314**:1053–66.

11. Xia LC, Ai D, Cram J, *et al.* Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* 2013;**29**:230–7.
12. Xia LC, Steele JA, Cram JA, *et al.* Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst Biol* 2011;**5**:S15.
13. Ruan Q, Dutta D, Schwalbach MS, *et al.* Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 2006;**22**: 2532–8.
14. He F, Zeng AP. In search of functional association from time-series microarray data based on the change trend and level of gene expression. *BMC Bioinformatics* 2006;**7**:69.
15. Ji L, Tan KL. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics* 2005;**21**:509–16.
16. ~~Xia LC, Ai D, Cram JA, *et al.* Statistical significance approximation in local trend analysis of high-throughput time-series data using the theory of Markov chains. *BMC Bioinformatics* 2015;**16**:301.~~
17. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
18. Vasily R, Alexandr S, Natalia Z, *et al.* A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* 2007;**69**:349–57.
19. Tachibana H, Inoue Y, Kanehisa T, Fukami Y. Local similarity in the amino acid sequence between the non-catalytic region of Rous sarcoma virus oncogene product p60v-src and intermediate filament proteins. *J Biochem* 2008;**104**:869–72.
20. Raptis A. Local similarity transformations for the boundary layer flow through a homogeneous porous medium by the presence of heat transfer. *Int Commun Heat Mass* 2000;**27**:739–43.
21. Lai Y, Wu B, Chen L, Zhao H. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* 2004;**20**:3146–55.
22. Li KC, Liu CT, Sun W, *et al.* A system for enhancing genome-wide coexpression dynamics study. *Proc Natl Acad Sci U S A* 2004;**101**: 15561–6.
23. Li X, Rao S, Jiang W, *et al.* Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics* 2006;**7**:26.
24. Wang L, Liu S, Ding Y, *et al.* Meta-analytic framework for liquid association. *Bioinformatics* 2017;**33**:2140–7.
25. Reshef DN, Reshef YA, Finucane HK, *et al.* Detecting novel associations in large data sets. *Science* 2011;**334**:1518–24.
26. Ai D, Li X, Liu G, *et al.* Constructing the microbial association network from large-scale time series data using granger causality. *Genes* 2019;**10**:216.
27. Ai D, Li X, Pan H, *et al.* Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics* 2019;**20**:185.
28. Weiss S, Van Treuren W, Lozupone C, *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 2016;**10**:1669–81.
29. Matchado MS, Lauber M, Reitmeier S, *et al.* Network analysis methods for studying microbial communities: a mini review. *Comput Struct Biotechnol J* 2021;**19**:2687–98.
30. Durno WE, Hanson NW, Konwar KM, Hallam SJ. Expanding the boundaries of local similarity analysis. *BMC Genomics* 2013;**14**(Suppl 1):S3.
31. Zhang F, Shan A, Luan Y. A novel method to accurately calculate statistical significance of local similarity analysis for high-throughput time series. *Stat Appl Genet Mol Biol* 2018; **17**:20180019.
32. Zhang F, Sun F, Luan Y. Statistical significance approximation for local similarity analysis of dependent time series data. *BMC Bioinformatics* 2019;**20**:53.
33. Wang YXR, Liu K, Theusch E, *et al.* Generalized correlation measure using count statistics for gene expression data with ordered samples. *Bioinformatics* 2018;**34**:617–24.
34. Tackmann J, Matias Rodrigues JF, von Mering C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Syst* 2019;**9**: 286–296.e8.
35. Liu L, Yang J, Lv H, Yu Z. Synchronous dynamics and correlations between bacteria and phytoplankton in a subtropical drinking water reservoir. *FEMS Microbiol Ecol* 2014;**90**:126–38.
36. Thiriet-Rupert S, Carrier G, Trottier C, *et al.* Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of *Tisochrysis lutea*. *Algal Res* 2018;**30**:59–72.
37. Lee YY, Lee SY, Lee SD, Cho KS. Seasonal dynamics of bacterial community structure in diesel oil-contaminated soil cultivated with tall fescue (*Festuca arundinacea*). *Int J Environ Res Public Health* 2022;**19**:4629.
38. Parada AE, Fuhrman JA. Marine archaeal dynamics and interactions with the microbial community over 5 years from surface to seafloor. *ISME J* 2017;**11**:2510–25.
39. Jones AC, Hambright KD, Caron DA. Ecological patterns among bacteria and microbial eukaryotes derived from network analyses in a low-salinity Lake. *Microb Ecol* 2018;**75**: 917–29.
40. Liang Z, Xu G, Shi J, *et al.* Sludge digestibility and functionally active microorganisms in methanogenic sludge digesters revealed by E. Coli-fed digestion and microbial source tracking. *Environ Res* 2021;**193**:110539.
41. Needham DM, Sachdeva R, Fuhrman JA. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* 2017;**11**: 1614–29.
42. Needham DM, Fuhrman JA. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 2016;**1**:16005.
43. Roux S, Chan LK, Egan R, *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat Commun* 2017;**8**:858.
44. Wang H, Sangwan N, Li HY, *et al.* The antibiotic resistome of swine manure is significantly altered by association with the Musca domestica larvae gut microbiome. *ISME J* 2017;**11**: 100–11.
45. Posch T, Eugster B, Pomati F, *et al.* Network of interactions between ciliates and phytoplankton during spring. *Front Microbiol* 2015;**6**:1289.
46. Džunková M, Martinez-Martinez D, Gardlík R, *et al.* Oxidative stress in the oral cavity is driven by individual-specific bacterial communities. *NPJ Biofilms and Microbiomes* 2018;**4**:29.
47. Wang Y, Ye J, Ju F, *et al.* Successional dynamics and alternative stable states in a saline activated sludge microbial community over 9 years. *Microbiome* 2021;**9**:199.
48. Shan A, Zhang F, Luan Y. Efficient approximation of statistical significance in local trend analysis of dependent time series. *Front Genet* 2022;**13**:729011.
49. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 2000;**97**:9834–9.

50. Nguyen TT, Almon RR, DuBois DC, *et al.* Importance of replication in analyzing time-series gene expression data: corticosteroid dynamics and circadian patterns in rat liver. *BMC Bioinformatics* 2010;**11**:279.
51. Zhu D, Li Y, Li H. Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data. *Bioinformatics* 2007;**23**:2298–305.
52. Yao J, Chang C, Salmi ML, *et al.* Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. *BMC Bioinformatics* 2008;**9**:288.
53. Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Stat Med* 2000;**19**: 1793–819.
54. Leroy AM, Rousseeuw PJ. *Robust regression and outlier detection*. John Wiley & Sons, 2005.
55. Hoaglin DC, Mosteller F, Tukey JW. *Understanding robust and exploratory data analysis*. Wiley series in probability and mathematical statistics, 1983.
56. Venables WN, Ripley BD. *Modern applied statistics with S*, Fourth edn. New York, NY: Springer-Verlag, 2002.
57. Sherman M, Speed FM, Jr, Speed FM. Analysis of tidal data via the blockwise bootstrap. *J Appl Stat* 1998;**25**:333–40.
58. Feller W. The asymptotic distribution of the range of sums of independent random variables. *Ann Math Statist* 1951;**22**: 427–32.
59. Andrews D. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 1991;**59**: 817–58.
60. Seekatz AM, Panda A, Rasko DA, *et al.* Differential response of the Cynomolgus macaque gut microbiota to Shigella infection. *PloS One* 2013;**8**:e64212.
61. Sun J, Liao XP, D'Souza AW, *et al.* Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. *Nat Commun* 2020;**11**:1427.
62. Zheng W, Huyan J, Tian Z, *et al.* Clinical class 1 integron-integrase gene - a promising indicator to monitor the abundance and elimination of antibiotic resistance genes in an urban wastewater treatment plant. *Environ Int* 2020;**135**: 105372.
63. Copeland E, Leonard K, Carney R, *et al.* Chronic Rhinosinusitis: potential role of microbial Dysbiosis and recommendations for sampling sites. *Front Cell Infect Microbio* 2018;**8**:57.
64. Xu J, Wang H, Xu R, *et al.* The diurnal fluctuation of colonic antibiotic resistome is correlated with nutrient substrates in a pig model. *Sci Total Environ* 2023;**891**:164692.
65. Jiang CL, Jin WZ, Tao XH, *et al.* Black soldier fly larvae (*Hermetia illucens*) strengthen the metabolic function of food waste biodegradation by gut microbiome. *J Microbial Biotechnol* 2019;**12**: 528–43.
66. Simons AL, Churches N, Nuzhdin S. High turnover of faecal microbiome from algal feedstock experimental manipulations in the Pacific oyster (Crassostrea gigas). *J Microbial Biotechnol* 2018;**11**:848–58.
67. Garcia J, Gannett M, Wei L, *et al.* Selection pressure on the rhizosphere microbiome can alter nitrogen use efficiency and seed yield in *Brassica rapa*. *Commun Biol* 2022;**5**:959.
68. Ki BM, Ryu HW, Cho KS. Extended local similarity analysis (eLSA) reveals unique associations between bacterial community structure and odor emission during pig carcasses decomposition. *J Environ Sci Health A Tox Hazard Subst Environ Eng* 2018;**53**: 718–27.
69. Pollet T, Berdjeb L, Garnier C, *et al.* Prokaryotic community successions and interactions in marine biofilms: the key role of *Flavobacteriia*. *FEMS Microbiol Ecol* 2018;**94**.
70. Chow CE, Sachdeva R, Cram JA, *et al.* Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California bight. *ISME J* 2013;**7**: 2259–73.
71. Ju F, Zhang T. Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant. *ISME J* 2015;**9**:683–95.
72. Kankan Z, Haodan Y, Ran X, *et al.* The only constant is change: endogenous circadian rhythms of soil microbial activities. *Soil Biol Biochem* 2022;**173**:108805.
73. Lee YY, Seo Y, Ha M, *et al.* Evaluation of rhizoremediation and methane emission in diesel-contaminated soil cultivated with tall fescue (Festuca arundinacea). *Environ Res* 2021;**194**: 110606.
74. Thomas F, Cébron A. Short-term rhizosphere effect on available carbon sources, Phenanthrene degradation, and active microbiome in an aged-contaminated industrial soil. *Front Microbiol* 2016;**7**:92.
75. Lee YY, Lee SY, Cho KS. Phytoremediation and bacterial community dynamics of diesel-and heavy metal-contaminated soil: long-term monitoring on a pilot scale. *Int Biodeter Biodegr* 2023;**183**:105642.
76. Lee YY, Choi H, Cho KS. Effects of carbon source, C/N ratio, nitrate, temperature, and pH on N2O emission and functional denitrifying genes during heterotrophic denitrification. *J Environ Sci Health A Tox Hazard Subst Environ Eng* 2019;**54**: 16–29.
77. Fletcher-Hoppe C, Yeh YC, Raut Y, *et al.* Symbiotic UCYN-A strains co-occurred with El Niño, relaxed upwelling, and varied eukaryotes over 10 years off Southern California. *ISME COMMUN* 2023;**3**:63.
78. Kwon JH, Park HJ, Lee YY, Cho KS. Evaluation of denitrification performance and bacterial community of a sequencing batch reactor under intermittent aeration. *J Environ Sci Health A Tox Hazard Subst Environ Eng* 2020;**55**:179–92.
79. Carini P, Delgado-Baquerizo M, Hinckley ES, *et al.* Effects of spatial variability and relic DNA removal on the detection of temporal dynamics in soil microbial communities. *MBio* 2020;**11**:e02776–19.
80. Kim TG, Yun J, Cho KS. The close relation between *Lactococcus* and *Methanosaeta* is a keystone for stable methane production from molasses wastewater in a UASB reactor. *Appl Microbiol Biotechnol* 2015;**99**:8271–83.
81. Lee YY, Kim TG, Cho KS. Effects of proton exchange membrane on the performance and microbial community composition of air-cathode microbial fuel cells. *J Biotechnol* 2015;**211**:130–7.
82. Steffen K, Indraningra AAG, Erngren I, *et al.* Oceanographic setting influences the prokaryotic community and metabolome in deep-sea sponges. *Sci Rep* 2022;**12**:3356.
83. Jang J, Park J, Hwang CY, *et al.* Abundance and diversity of antibiotic resistance genes and bacterial communities in the western Pacific and southern oceans. *Sci Total Environ* 2022;**822**:153360.
84. Zhuang Y, Chai J, Cui K, *et al.* Longitudinal investigation of the gut microbiota in goat kids from birth to Postweaning. *Microorganisms* 2020;**8**:1111.
85. Bergk Pinto B, Maccario L, Dommergue A, *et al.* Do organic substrates drive microbial community interactions in Arctic snow? *Front Microbiol* 2019;**10**:2492.

86. Auladell A, Sánchez P, Sánchez O, *et al.* Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria. *ISME J* 2019;**13**:1975–87.
87. Daudin JJ, Etienne MP, Vallois P. Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stoch Proc Appl* 2003;**107**:1–28.
88. He F, Chen H, Probst-Kepper M, *et al.* PLAU inferred from a correlation network is critical for suppressor function of regulatory T cells. *Mol Syst Biol* 2012;**8**:624.
89. Gonçalves PJ, Aires SR, Francisco PA, Madeira SC. Regulatory snapshots: integrative mining of regulatory modules from expression time series and regulatory networks. *PloS One* 2017;**7**:e35977.
90. Sudhakar P, Reck M, Wang W, *et al.* Construction and verification of the transcriptional regulatory response network of Streptococcus mutans upon treatment with the biofilm inhibitor carolacton. *BMC Genomics* 2014;**15**:362.