BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 230–237
doi:10.1093/bioinformatics/bts668

Systems biology

Advance Access publication November 23, 2012

Efficient statistical significance approximation for local similarity analysis of high-throughput time series data

Li C. Xia[1,†], Dongmei Ai[1,2,†], Jacob Cram[3], Jed A. Fuhrman[3] and Fengzhu Sun[1,4,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-2910, USA, [2]Department of Information and Computational Sciences, University of Science and Technology Beijing, Beijing, 100083, China, [3]Marine and Environmental Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-0371, USA and [4]TNLIST/Department of Automation, Tsinghua University, Beijing, China.

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Local similarity analysis of biological time series data helps elucidate the varying dynamics of biological systems. However, its applications to large scale high-throughput data are limited by slow permutation procedures for statistical significance evaluation.

Results: We developed a theoretical approach to approximate the statistical significance of local similarity analysis based on the approximate tail distribution of the maximum partial sum of independent identically distributed (i.i.d.) random variables. Simulations show that the derived formula approximates the tail distribution reasonably well (starting at time points >10 with no delay and >20 with delay) and provides *P*-values comparable with those from permutations. The new approach enables efficient calculation of statistical significance for pairwise local similarity analysis, making possible all-to-all local association studies otherwise prohibitive. As a demonstration, local similarity analysis of human microbiome time series shows that core operational taxonomic units (OTUs) are highly synergetic and some of the associations are body-site specific across samples.

Availability: The new approach is implemented in our eLSA package, which now provides pipelines for faster local similarity analysis of time series data. The tool is freely available from eLSA's website: http://meta.usc.edu/softs/lsa.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: fsun@usc.edu

Received on June 1, 2012; revised on September 12, 2012; accepted on November 12, 2012

1 INTRODUCTION

Understanding how genes regulate each other and when the regulations are active is an important problem in molecular biological research. Similarly, in ecological studies, it is important to understand how different organisms and environmental factors, such as food resources, temperature, etc., regulate each other to affect the whole community. For generality, we will refer to either genes in gene regulation studies or organisms or environmental factors in ecological studies as *factors*. Time series data can give significant insights about the regulatory relationships among different factors. Many computational or statistical approaches have been developed to cluster the genes into different groups so that the expression profiles of genes in each cluster are highly correlated (Androulakis *et al.*, 2007; Bar-Joseph, 2004). Most of these methods consider the correlation of expression patterns across the entire time interval of interest. For many gene regulation relationships, the regulation may be active in certain subintervals. Methods based on the global associations of the gene expression profiles may fail to detect these relationships.

Several local association-based methods have been developed to address this problem (Qian *et al.*, 2001; Balasubramaniyan *et al.*, 2005; Ji and Tan, 2005; He and Zeng, 2006). Borrowing the idea from local alignment for molecular sequences, Qian *et al.* (2001) proposed to identify local and potential time-delayed (-lagged) associations between gene expression profiles. Here, local indicates the two factors are only associated within some time subinterval, and time-delayed indicates there is time shift in the associated profiles. The strength of local association is measured by local similarity (LS) score and the statistical significance of LS score is evaluated by a large number of permutations. The authors showed that such analysis can identify associated pairs that are not detectable through global analysis. Ruan *et al.* (2006) used a similar approach to study local associations of microbial organisms in the ocean over a 4-year period, and this approach has been used in several other recent ecological studies (Beman *et al.*, 2011; Chaffron *et al.*, 2010; Gilbert *et al.*, 2011; Shade *et al.*, 2010; Steele *et al.*, 2011). Xia *et al.* (2011b) recently extended the approach to deal with replicated time series where not only statistical significance of LS score can be evaluated, but also a bootstrap confidence interval can be obtained.

One of the major limitations of the local similarity analysis is the time-consuming permutation procedure used to evaluate the statistical significance (*P*-value) of the LS score. When a large number of (G) genes are considered, $G(G-1)/2$ gene pairs need to be evaluated. For a type I error α, in order to adjust for multiple testing, the Bonferroni corrected threshold is $2\alpha/(G(G-1))$. For $G = 5000$, the threshold is 4×10^{-9} when $\alpha = 0.05$, which will need over 2.5×10^8 permutations that are computationally prohibitive. Although in practice false discovery rate (*Q*-value) is used to correct for the multiple comparison problem, still, thousands of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

permutations are needed for estimating *P*-values that were used for *Q*-value calculation. A fast and efficient theoretical approximation for the statistical significance of the LS score is urgently needed. Recently, Durno *et al.* (2013) provided an upper bound for the p-value corresponding to a LS score. However, the upper bound is not tight.

We make the following contributions in this article. First, Feller (1951) developed an approximation theory for the range of partial sums for independent identically distributed (i.i.d.) random variables with mean zero, and Daudin *et al.* (2003) extended the results to Markovian random variables. Although the theory developed by Feller (1951) was later further extended by others, it has not been applied to the field of computational biology. We are the first to use Feller's theory to approximate the distribution of LS scores. On the other hand, the theory corresponding to the scenario that the expectation of each random variable is negative was used by Karlin *et al.* (1990) and Karlin and Altschul (1993) to derive the statistical significance for local sequence alignment. Such a development was crucial for the wide use of BLAST (Altschul *et al.*, 1990) in computational biological research.

Second, the theory by Feller (1951) and others are valid only when the number of summands (time points) is large. However, it is not clear how large the number of time points should be so that the approximation is reasonable for the LS score. We show through simulations that the approximation from Feller (1951) is reasonable as long as the number of time points is at least 10 for LS score without time delay. In addition, we show how we can adapt Feller's approximation to calculate statistical significance of LS score with time delays. Simulations showed that the resulting approximation is appropriate when the number of time points is above 20. Finally, we applied the developed theory to the analysis of real datasets from gene expression profiles to metagenomics communities, where interesting biological results were obtained.

The organization of the article is as follows. In the 'Methods' section, we provide the theoretical bases for deriving the approximate tail probability that the LS score is above a threshold. In the 'Results' section, we use simulations to study the number of data points n needed for the theoretical approximation to be valid. We also use the theoretical formula to study three real datasets arising from different high-throughput experiments: microarray, molecular finger printing and NGS tag-sequencing. The article concludes with some discussion on further applications and future research directions.

2 METHODS

Consider time series data for two factors with levels $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_n$, respectively. The first step is to normalize the expression levels of each time series so that they can be regarded as normally distributed. Without loss of generality, we assume that they are already normally distributed. Second, dynamic programming algorithm is used to find intervals $I = [i, i+l-1]$ and $J = [j, j+l-1]$ of the same length l such that the absolute value of $S = \sum_{k=0}^{l-1} X_{i+k} Y_{j+k}$ is maximized, which is referred to as local similarity (LS) score (Qian *et al.*, 2001; Ruan *et al.*, 2006; Xia *et al.*, 2011b). Here the starting positions of the subintervals i and j, and the length of the intervals l are not pre-specified and are all derived from the data. In most practical problems, investigators may only be interested in local associations with short delays, for example, the starting positions of the intervals in the two time series i and j are at most D units apart, i.e. $|i-j| \le D$. We denote the LS score with time delay at most D units by $LS(D)$.

In the third step, statistical significance for the LS score corresponding to the null hypothesis that the two factors are not associated is approximated by permuting one of the time series data many times and calculating the fraction of times that the LS score for the permuted data is higher than that for the real data (Qian *et al.*, 2001; Ruan *et al.*, 2006). With such a permutation approach for *P*-value, the authors implicitly assumed that the observations for the samples at the different time points are independent under the null model. However, in many practical problems, in particular, time series data, the observations for each factor may depend on each other and the permutation-based approach may not work well. Another drawback of the permutation-based approach is that computational time scales linearly with the inverse of the *P*-value precision and is computationally expensive for large dataset of long series. Here, we provide theoretical formulas to approximate the *P*-value overcoming both problems.

2.1 Maximum absolute partial sums of i.i.d. and Markovian random variables

To derive theoretical formulas to approximate the *P*-value related to the local similarity score, we resort to classical theoretical studies on the range of partial sums for i.i.d. random variables with zero mean (Feller, 1951). The results from such studies when the expectation of the random variables is negative played key roles in the derivation of statistical significance for local sequence alignment, e.g. BLAST (Altschul *et al.*, 1990), which forms a milestone in the field of computational biology (Karlin *et al.*, 1990; Karlin and Altschul, 1993). On the other hand, the theoretical results on the approximate distributions when the mean is zero have not been used in the computational biology community.

Based on these previous theoretical studies, we present some theoretical results regarding the range of partial sums for either i.i.d. or Markovian random variables. Feller (1951) studied the approximate distribution of the range of the sum of n random variables with mean 0. Let Z_i be i.i.d. random variables such that $E(Z_i) = 0$ and $Var(Z_i) = \sigma^2$. Let $S_n = Z_1 + Z_2 + \cdots + Z_n$, $M_n = \max\{0, S_1, S_2, \cdots, S_n\}$, and $m_n = \min\{0, S_1, S_2, \cdots, S_n\}$. The range is defined as $R_n = M_n - m_n$. It is shown in Feller (1951): $E(R_n/\sigma) = 2\sqrt{2n/\pi}$, $\mathrm{Var}(R_n/\sigma) = 4n(\log(2) - 2/\pi)$.

Using the theory of Bachelier–Wiener processes, Feller (1951) approximated the density function of R_n/σ by (equations 3.7 and 3.8 in that article) $\delta(n; r)$,

$$\delta(n; r) = \sqrt{\frac{2}{\pi}} r^{-1} L'(r/(2\sqrt{n})), \tag{1}$$

where,

$$L(z) = \sqrt{2\pi} z^{-1} \sum_{k=0}^{\infty} \exp\left(-(2k+1)^2\pi^2/8z^2\right). \tag{2}$$

Thus,

$$\begin{aligned} P(R_n/(\sigma\sqrt{n}) \ge x) &= \int_{\sqrt{n}x}^{\infty} \sqrt{\frac{2}{\pi}} r^{-1} L'\left(\frac{r}{2\sqrt{n}}\right) dr \\ &= 1 - 8\sum_{k=0}^{\infty} \left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right) \exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \end{aligned} \tag{3}$$

Since the summation in equation 3 is an infinite sum, we need to decide when to stop for numerical approximations. To achieve this objective, we next give an upper bound for the tail in equation (3).

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

This upper bound can be used to determine when we stop the summation in equation (3) for practical calculations. Note $\exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right) < \exp\left(-\frac{(2k+1)\pi^2}{2x^2}\right)$, for $k>0$. Thus, for any $K>0$ such that $(2K+1)\pi > x$, we have,

$$\begin{aligned}&\sum_{k=K}^{\infty}\left(\frac{1}{x^2}+\frac{1}{(2k+1)^2\pi^2}\right)\exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right)\\&<\frac{2}{x^2}\sum_{k=K}^{\infty}\exp\left(-\frac{(2k+1)\pi^2}{2x^2}\right)\\&=\frac{2\exp\left(-\frac{(2K+1)\pi^2}{2x^2}\right)}{x^2\left(1-\exp\left(-\frac{2\pi^2}{2x^2}\right)\right)}.\end{aligned} \tag{4}$$

Thus, for an error threshold β, we can choose K so that

$$\frac{16\exp\left(-\frac{(2K+1)\pi^2}{2x^2}\right)}{x^2\left(1-\exp\left(-\frac{2\pi^2}{2x^2}\right)\right)} \le \beta. \tag{5}$$

Then we approximate $P(R_n/(\sigma\sqrt{n}) \ge x)$ by

$$1-8\sum_{k=0}^{K-1}\left(\frac{1}{x^2}+\frac{1}{(2k+1)^2\pi^2}\right)\exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \tag{6}$$

Daudin *et al.* (2003) studied the distribution of the maximum partial sum of an aperiodic Markov chain taking values on a finite subset of the real line, i.e $H_n = \max_{0\le i<j\le n}; (S_j - S_i)$. Let v be the stationary distribution of the Markov chain Z_n, $n = 0, 1, 2, \cdots$ with $E_v(Z_1) = 0$ and $\sigma^2 = E_v(Z_1^2) + 2\sum_{k=2}^{\infty} E_v(Z_1 Z_k)$. It was shown in Daudin *et al.* (2003) that

$$\lim_{n\to\infty} P\left(\frac{H_n}{\sigma\sqrt{n}} \le x\right) = \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right). \tag{7}$$

Etienne and Vallois (2004) provided an upper bound of $C\sqrt{\log(n)/n}$ for the approximation in equation (7).

Similarly, we can define $L_n = -\min_{0\le i<j\le n}; (S_j - S_i)$. Since $E(Z_i) = 0$, L_n has the same limiting distribution as H_n. It can also be seen easily that $R_n = \max(H_n, L_n)$. When x is large, the probability of $\{H_n > x\} \cap \{L_n > x\}$ will be small and

$$\begin{aligned}&P(R_n/(\sigma\sqrt{n}) \ge x) = P(\max(H_n, L_n)/(\sigma\sqrt{n}) \ge x)\\&= P(\{H_n/(\sigma\sqrt{n}) \ge x\}\bigcup\{L_n/(\sigma\sqrt{n}) \ge x\})\\&\approx P(H_n/(\sigma\sqrt{n}) \ge x) + P(L_n/(\sigma\sqrt{n}) \ge x)\\&\approx 2P(H_n/(\sigma\sqrt{n}) \ge x)\\&\approx 2\left(1-\frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right)\right), x \ge 0.\end{aligned} \tag{8}$$

The approximation works well when $x \ge 2$. However, when x is small, the approximation does not work well and actually the above quantity can be larger than 1.

2.2 Statistical significance for local similarity scores

We next use the theory outlined in subsection 2.1 to approximate the statistical significance in local similarity analysis. For time series data of two factors $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_n$ that have been normalized as in Qian *et al.* (2001) and Ruan *et al.* (2006), we use the dynamic programming algorithm to calculate the LS score with maximum delay D denoted as s_D. Corresponding to the null hypothesis that the two time series data are not related, the statistical significance is given by $P-\text{value} = P(LS(D) \ge s_D)$, where $LS(D) = \max_{i,j,l;|i-j|\le D} |\sum_{k=0}^{l-1} X_{i+k} Y_{j+k}|$. First consider the case that $D = 0$. Let $Z_i = X_i Y_i, i = 1, 2, \cdots, n$. Assuming that both X_i and Y_i are independent standard normally distributed, we have $E(Z_i) = 0$ and $\sigma^2 = E(Z_i^2) = 1$. Therefore, we can directly use the theory developed above in equations (6) to calculate the P-value.

Next let us assume $D>0$. Let $S_n^{(d)}$ be the LS score with no time delay for the pair of series ($d = 0, \pm 1, \pm 2, \cdots, \pm D$)

x_1	x_2	x_3	·····	x_{n-2}	x_{n-1}	x_n
y_{1+d}	y_{2+d}	y_{3+d}	·····	x_{n-2+d}	x_{n-1+d}	x_{n+d}

where we consider the data as missing when the subscript is outside the range $[1, n]$ and the pair is not considered when the LS score is calculated. When n is sufficiently large, $S_n^{(d)}$ for $d = 0, \pm 1, \pm 2, \cdots, \pm D$ can be considered as approximately identically distributed because $S_n^{(d)}$ is the LS score for $n - d$ pairs of i.i.d. normal random variables. The tail distribution function of $S_n^{(d)}/(\sigma\sqrt{n})$ can be approximated by equation (6). Note $LS(D) = \max_{d=-D}^{D} S_n^{(d)}$. To derive an approximate cumulative distribution function of $LS(D)$, we pretend that $S_n^{(d)}, d = 0, \pm 1, \pm 2, \cdots, \pm D$ are independent although they are not. Then,

$$\begin{aligned}&P(LS(D)/(\sigma\sqrt{n}) \le x)\\&= \prod_{d=-D}^{D} P(S_n^{(d)}/(\sigma\sqrt{n}) \le x)(\text{use independence assumption})\\&= 8^{2D+1}\left(\sum_{k=1}^{\infty}\left(\frac{1}{x^2}+\frac{1}{(2k-1)^2\pi^2}\right)\exp\left(-\frac{(2k-1)^2\pi^2}{2x^2}\right)\right)^{2D+1}.\end{aligned} \tag{9}$$

Thus, the tail probability of $LS(D)$ can be approximated by

$$\begin{aligned}&\mathcal{L}(x) = P(LS(D)/(\sigma\sqrt{n}) \ge x)\\&\approx 1-8^{2D+1}\left(\sum_{k=1}^{\infty}\left(\frac{1}{x^2}+\frac{1}{(2k-1)^2\pi^2}\right)\exp\left(-\frac{(2k-1)^2\pi^2}{2x^2}\right)\right)^{2D+1}.\end{aligned} \tag{10}$$

In the following numerical calculations shown in Figure 1, we only use the first 200 terms to approximate the infinite series in the above equation.

From equation (10), we can obtain the approximate density function of $R_n^{(D)}/(\sigma\sqrt{n})$ by

$$\begin{aligned}&f_D(x)\\&=\frac{(2D+1)8^{2D+1}}{x^3}\left(\sum_{k=1}^{\infty}\left(\frac{1}{x^2}+\frac{1}{(2k-1)^2\pi^2}\right)\exp\left(-\frac{(2k-1)^2\pi^2}{2x^2}\right)\right)^{2D}\\&\times\sum_{k=1}^{\infty}\left(\frac{(2k-1)^2\pi^2}{x^2}-1\right)\exp\left(-\frac{(2k-1)^2\pi^2}{2x^2}\right).\end{aligned} \tag{11}$$

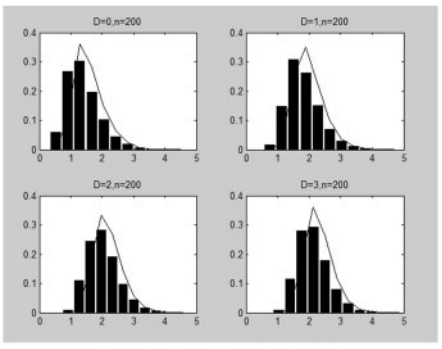

Fig. 1. The histogram of local similarity scores $LS(D)/\sqrt{n}$ for $n = 200$ and $D = 0, 1, 2, 3$ together with the theoretical approximate density function given in equation (11)

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

2.3 Dealing with replicates

To reduce the effect of biological and/or technical variation on the LSA results, replicate experiments are frequently carried out (Quinn and Keough, 2002). An extended LSA (eLSA) approach was developed for time series data with replicates (Xia *et al.*, 2011b). First, for each sample the replicate data at each time point were summarized by a function, for example, the average over the replicates. Second, the local similarity score is calculated using the averages for the sample pairs. Third, statistical significance for testing the hypothesis that the two sequences are related is obtained by randomly shuffling the data along the different time points. Finally, the bootstrap confidence interval for the LS score is obtained by bootstrapping the data at each time point by sampling from the observed data with replacement.

With the theory developed above, we can significantly speed up the process of evaluating the statistical significance of the LS score in the third step. Let $X^{(m)} = (X_1^{(m)}, \cdots, X_n^{(m)})$ and $Y^{(m)} = (Y_1^{(m)}, \cdots, Y_n^{(m)})$ be the m-th replicate for the time series data, $m = 1, 2, \cdots, M$. The essence of eLSA is to calculate the local similarity score of $U_i = F(X_i^{(1)}, \cdots, X_i^{(M)})$ and $V_i = F(Y_i^{(1)}, \cdots, Y_i^{(M)})$, where $F(\cdot)$ is the summarizing function. Then by replacing X and Y in non-replicated case with U and V, respectively, similar approaches can be used to obtain the P-value. In particular, if $U_i = F(X_i^{(1)}, \cdots, X_i^{(M)}) = \bar{X}_i = \sum_{m=1}^{M} X_i^{(m)}/M$ and all the $X_i^{(m)}$ are standard normal, then $\mathrm{Var}(U_i) = 1/M$. Similarly, we have $\mathrm{Var}(V_i) = 1/M$ assuming that each of the $X_i^{(m)}$ is already standard normal. Thus, $\sigma^2 = \mathrm{Var}(U_i V_i) = 1/M^2$. Let the local similarity score with time delay at most D for the real data be s_D. Then the P-value is calculated by

$$\begin{aligned} &P(LS(D) \geq s_D) \\ &= P\left(\frac{M \times LS(D)}{\sqrt{n}} \geq M \times s_D/\sqrt{n}\right) \\ &= \mathcal{L}(M \times s_D/\sqrt{n}), \end{aligned} \tag{12}$$

where the function $\mathcal{L}$ is defined in equation (10).

2.4 Data normalization

In reality, normality may not be satisfied by the raw data. Through normalization, the normality of the data can be ensured for subsequent analysis. To accommodate possible nonlinear associations and the variation of scales within the raw data, we apply the following approach to normalize the raw data before any LS score calculations as described in Li (2002). We use x_i to denote the original raw data of the i-th time spot of a factor X. First, we take r_k = rank of x_k in $\{x_1, x_2, ..., x_n\}$. Then, we take $s_k = \Phi^{-1}(r_k/n+1)$, where Φ is the cumulative distribution function of the standard normal distribution.

In case of small n, we find that the above transformed data $S = s_{[1:n]}$ do not necessarily follow a standard normal distribution closely. When the variance is not 1 and mean is not zero, it will cause the LS scores calculated to be smaller than that expected from the theory and can lead to unexpected high P-values. To overcome this difficulty, we further scale and shift $S = s_{[1:n]}$ using the Z-score transformation, such that $z_i = s_i - \bar{S}/\sigma_S$. We will take $Z = z_{[1:n]}$ as the standardized normalization of X.

2.5 Simulation studies and applications to real datasets

In deriving the approximate P-values for local similarity analysis in subsection 2.2, we made several simplifying assumptions, whose effects on the accuracy of the approximations were evaluated by simulations. We first study the accuracy of the approximation for the tail probability of $R_n/(\sigma\sqrt{n})$ in equation (10) using simulations for local similarity analysis. Firstly, for given number of time points n, we generate n pairs of i.i.d. standard normal random variables (X_i, Y_i), $i = 1, 2, \cdots, n$, where X_i and Y_i are independent. Secondly, the dynamic programming algorithm implemented in eLSA (Xia *et al.*, 2011b) package is used to calculate the local similarity score with at most time delay D, $LS(D)$. Thirdly, we repeat the first two steps 10,000 times and obtain the empirical distribution of $LS(D)/(\sigma\sqrt{n})$. We compare the empirical distributions with the theoretical approximation given in equation (10) with $\sigma = 1$.

We then apply our method to analyze three real datasets. The first one is a microarray yeast gene expression dataset, synchronized by the cdc-15 gene, from Spellman *et al.* (1998) (referred to as 'CDC'). The second one is an ARISA molecular finger printing microbial ecology dataset from San Pedro Ocean Time Series in Steele *et al.* (2011) (referred to as 'SPOT'). The third one is a 16S RNA tag-sequencing dataset from the 'Moving Pictures of Human' sampling of human symbiotic microbial communities from Caporaso *et al.* (2011) (referred to as 'MPH'). We apply local similarity analysis to re-analyze the first two datasets and compare the theoretical and permutation p-values. We are the first to analyze the third dataset using local similarity analysis.

3 RESULTS

3.1 Simulations

The approximate P-value for the local similarity score given in subsection 2.2 is only applicable when the P-value is small and the number of time points is large. Thus, it is important to know the range of applicability for the approximation. Table 1 gives the theoretical tail probability based on equation (10) (2nd column) and the simulated probability $P(LS(0)/\sqrt{n} \geq x)$ (3rd to 9th columns) for different number of time points when $D = 0$. It can be seen that the theoretical tail probability is very close to the simulated probability when the theoretical P-value is less than 0.01. The approximation is even reasonable when the number of time points is just 10. In general, the theoretical tail probability is slightly larger than the simulated values when $D = 0$ (Table 1 and Supplementary Table S1). When $D = 1, 2, 3$, the theoretical approximation is close to the simulated tail probability when $n \geq 20$ and the theoretical P-value is less than 0.01 (for $D = 3$ see Table 2 and also see Supplementary Tables S2–S4 in Supplementary Results). Thus, if we use the theoretical approximate distribution to calculate the P-value, we will be slightly conservative in declaring significant associations.

For relatively small value of x, the theoretical approximation can be much larger than the simulated tail probability. One potential explanation is that $R_n^{(D)}/(\sigma\sqrt{n})$ is stochastically increasing with respect to n and that the theoretical approximation becomes closer to the simulated distribution of $LS(D)/(\sigma\sqrt{n})$ as n increases. We also tested if $P(LS(D)/(\sigma\sqrt{n}) \geq x) = 1 - (1 - P(R_n^{(0)}/(\sigma\sqrt{n}) \geq x))^{2D+1}$ is generally true using the simulated tail probabilities in Figure 1, and it can be clearly seen from Supplementary Tables S1–S4 that this relationship is indeed reasonable, indicating that $S_n^{(d)}, d = 0, \pm 1, \cdots \pm D$ can effectively be considered as independent.

In equation (11), we derive the approximate density function of $R_n^{(D)}/(\sigma\sqrt{n})$. We superimpose this approximate density function to the histograms of the simulated $LS(D)/(\sigma\sqrt{n})$ at $n = 200$ and $D = 0, 1, 2, 3$ in Figure 1. Several observations can be made from the figure. First, the values of $LS(D)/(\sigma\sqrt{n})$ increase as a function of D as expected. Second, the approximate theoretical

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

Table 1. Theoretical approximation for local similarity analysis *P*-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$

		Number of time points n						
x	Theory	10	20	30	40	60	80	100
2	0.1815	0.0848	0.0987	0.1062	0.1122	0.1201	0.1235	0.1290
2.2	0.1111	0.0541	0.0621	0.0645	0.0665	0.0699	0.0771	0.0767
2.4	0.0656	0.0341	0.0367	0.0392	0.0416	0.0411	0.0435	0.0457
2.6	0.0373	0.0223	0.0221	0.0252	0.0235	0.0232	0.0249	0.0261
2.8	0.0204	0.0147	0.0128	0.0154	0.0131	0.0129	0.0138	0.0163
3.0	0.0108	0.0093	0.0082	0.0088	0.0074	0.0069	0.0071	0.0090
3.2	0.0055	0.0056	0.0051	0.0038	0.0036	0.0030	0.0035	0.0054
3.4	0.0027	0.0033	0.0031	0.0017	0.0022	0.0009	0.0016	0.0027
3.6	0.0013	0.0019	0.0020	0.0011	0.0014	0.0004	0.0006	0.0012
3.8	0.0006	0.0007	0.0008	0.0006	0.0010	0.0002	0.0004	0.0009
4.0	0.0003	0.0004	0.0005	0.0003	0.0005	0.0000	0.0003	0.0004
4.2	0.0001	0.0002	0.0004	0.0002	0.0005	0.0000	0.0001	0.0002
4.4	0.0000	0.0001	0.0003	0.0001	0.0002	0.0000	0.0000	0.0001
4.6	0.0000	0.0000	0.0003	0.0001	0.0000	0.0000	0.0000	0.0001
4.8	0.0000	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000

Note: The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 9th columns. $D = 0$.

density function is slightly lower than the simulated frequency when x is lower than the mode of the theoretical distribution and is slightly higher than the simulated frequency when x is larger than the mode of the theoretical distribution, thus the tail probability based on the theoretical approximation is slightly higher than the simulated value.

We next see how *P*-values (P_{theo}) derived from theoretical approximation compare with that of permutation (P_{perm}) given the same data (Supplementary Figure S1) in simulation. Starting from $D = 0$ and $n = 20$, points in scatter plots become concentrated on the diagonal line (where $P_{perm} = P_{theo}$) and they become more aligned as n increases. This indicates an increasing rate of agreement between the theoretical and permutation *P*-values, representing their reasonable approximation to the null distribution in spite of the inherent variance associated with the permutation procedures. The same is true with $D > 0$, and the theoretical approximation become significantly closer to the permutation one as n increase. Though, when $D > 0$, the variation seems more substantial and close alignment only starts at $n = 30$. In summary, we can see that if we are interested in statistical significance given some type I error threshold, the theoretical approach shall provide results comparable with that from permutations starting from $n = 20$.

3.2 The CDC dataset

The CDC dataset consists of the expression profiles of 6177 genes at 24 time points. It is extremely time consuming to approximate the *P*-values for all the gene pairs using permutations. Thus, we only randomly selected 25 genes and estimated the *P*-value for each of the 300 gene pairs by permuting the original data 1000 times. We then compared P_{theo}s from our theoretical approximation to P_{perm}s from the permutation approach, shown in Figure 2. It can be seen from the figure that P_{theo} is highly positively correlated with P_{perm}, but P_{theo} is slightly higher than P_{perm}, indicating that it is conservative when we declare statistical significance using P_{theo}.

In particular, we compare the gene pairs declared as significant by either P_{theo} or P_{perm} for the type-I error threshold 0.05 in Supplementary Table S5. For all the situations considered, none of P_{theo} is less than 0.05 when $P_{perm} > 0.05$. Among the gene pairs with $P_{perm} \leq 0.05$, over half of them are declared as significant by P_{theo}. For the local similarity analysis, using $D = 0$, we have 233 (78%) out of 300 found to be non-significant by both theoretical approximation and permutations. Among the remaining, 48 (16%) are found significant by both methods, and in total 281 (94%) are in agreement. The results are similar with $D = 1,2,3$, with 262 (88%), 262 (88%) and 262 (88%) in agreement, respectively. Moreover, all-to-all pairwise analysis of the whole CDC dataset with $D = 3$ and permutation 1000 times cannot be completed in 100 hours on a 'Dell, PE1950, Xeon E5420, 2.5 GHz, 12010 MB RAM' computing node, while, using the theoretical approach, it finishes in 10 hours on the same node.

3.3 The SPOT dataset

The SPOT dataset consists of 10-year monthly (114 time points) sampled operational taxonomic unit (OTU) abundance data. As above, we selected 40 abundant OTUs from the SPOT dataset with the criteria 'the OTU occurs at least 20 times with minimum relative abundance 1% and has less than 10 missing values'. We summarize *P*-value comparison for local similarity analysis in Supplementary Table S6 and Figure 2.

With $D = 0$ and type-I error 0.05, we have 488 (63%) out of 780 found non-significant and 261 (33%) significant by both methods. In total, 685 (96%) are in agreement. All of the remaining 31 (4%) pairs are significant by P_{perm} but non-significant by P_{theo}. The results are similar with $D = 1,2,3$, where 733 (94%), 723 (93%) and 727 (93%) are in concordance, respectively. There

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

Table 2. Theoretical approximation for local similarity analysis *P*-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$

		The number of time points n						
x	Theory	10	20	30	40	60	80	100
2	0.7539	0.2509	0.3331	0.4103	0.4544	0.5120	0.5402	0.5571
2.2	0.5616	0.1731	0.2301	0.2772	0.3124	0.3533	0.3707	0.3917
2.4	0.3779	0.1177	0.1486	0.1772	0.2000	0.2293	0.2344	0.2539
2.6	0.2336	0.0785	0.0952	0.1122	0.1236	0.1366	0.1401	0.1578
2.8	0.1346	0.0513	0.0583	0.0679	0.0736	0.0767	0.0813	0.0918
3.0	0.0732	0.0320	0.0343	0.0379	0.0416	0.0443	0.0453	0.0534
3.2	0.0379	0.0199	0.0205	0.0207	0.0210	0.0237	0.0264	0.0278
3.4	0.0187	0.0123	0.0129	0.0116	0.0110	0.0124	0.0122	0.0142
3.6	0.0089	0.0083	0.0073	0.0055	0.0059	0.0058	0.0061	0.0076
3.8	0.0040	0.0047	0.0046	0.0030	0.0029	0.0029	0.0036	0.0048
4.0	0.0018	0.0028	0.0032	0.0013	0.0015	0.0015	0.0013	0.0024
4.2	0.0007	0.0019	0.0021	0.0004	0.0007	0.0010	0.0005	0.0008
4.4	0.0003	0.0009	0.0015	0.0002	0.0004	0.0007	0.0002	0.0002
4.6	0.0001	0.0006	0.0009	0.0002	0.0003	0.0003	0.0000	0.0001
4.8	0.0000	0.0003	0.0002	0.0000	0.0002	0.0000	0.0000	0.0001
5.0	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0001

Note: The theoretical approximate probability based on equation (10) with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{n} \geq x$ is given in the 3rd to the 9th columns. $D = 3$.

are about 6-7% associations significant by P_{perm} but non-significant by P_{theo}, showing that P_{theo} is more conservative. The agreement of the significant results between P_{theo} and P_{perm} for the SPOT dataset is better than that for the CDC dataset. This can be explained by the fact that the number of time points for the SPOT data (114) is much higher than that for the CDC dataset (24) and the approximation is better when the number of time points is large.

3.4 The MPH dataset

The MPH dataset consists of 130, 133 and 135 daily sequenced samples from feces, palm and tongue sites of a female ('F4'), and 332, 357 and 372 samples from a male ('M3'), respectively (Caporaso *et al.*, 2011). The genus level OTU abundance is used in our analysis. There are 335, 1295 and 373 unique OTUs from feces, palm and tongue sites of 'F4' and 'M3', respectively. With $D = 3$, we analysed the MPH dataset with local similarity analysis. Because of the intra-person variability (both time and site) of the human microbiota, one important step in analysing human microbiota datasets is to identify the core (persisting) group of microbes for a specific body site of a person. Based on the discussion in Caporaso *et al.* (2011), we consider core OTUs as those showing in at least 60% of samples from the same body site of one person. Using this criteria, we identified 45, 252 and 41 core OTUs for the feces, palm and tongue sites of 'F4', and 59, 269 and 56 core OTUs for the corresponding sites of 'M3', respectively.

Subsequent analysis showed these symbiotic core microbes are highly synergetic. We used local similarity analysis as our main approach and report significant local associations with *P*-value ≤ 0.05, *Q*-value ≤ 0.05 and Aligned Length $\geq 80\%$ time points (Xia *et al.*, 2011b). We found 194 significant associated pairs within the subset formed by the 45 core OTUs of the 'F4' feces samples and the intra association rate (the average degree

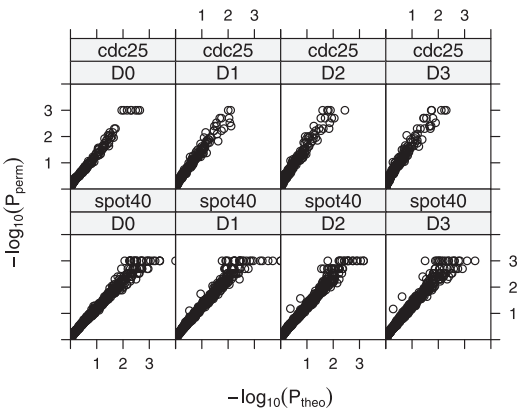

Fig. 2. The comparison of P_{theo} and P_{perm} for all-to-all pairwise local similarity analysis of 25 factors from the CDC dataset ('cdc25') and 40 factors from the SPOT dataset ('spot40'). Columns D0 to D3 are for $D = 0, 1, 2, 3$, respectively

divided by the number of OTUs minus one in an association subnetwork) is 20%. The rates are 32% and 20% for 'F4' tongue and palm samples, whereas 53%, 55% and 72% for 'M3' feces, tongue and palm samples, respectively. These percentages translate into a picture of high connectivity between these OTUs, supporting their role as crucial players in the corresponding microbiota.

However, are these site-specific significant local associations shared across individuals? We used Sorensen index Q_s to measure the similarity between significant set of local associations for any two samples (Sorensen, 1948). We consider only the OTUs common to the two samples. Suppose the subsets of significant local associations between their common OTUs are S_1 and S_2 for the two samples. Then the Sorensen index is defined as

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

$\frac{2|S_1 \cap S_2|}{|S_1|+|S_2|}$, which is two times the number of shared associations divided by the sum of number associations in each sample. The index ranges from 0 to 1 and is higher for community with more similar associations.

We found 'F4' and 'M3' shared 129 pairs of their significant local associations ($Q_s = 0.57$) in feces samples. These are mostly associations between the *Lachnospiraceae*, *Ruminococcaceae*, *Erysipelotrichaceae*, *Enterobacteriaceae* and *Veillonellaceae* family members. At tongue and right palm sites, 192 pairs ($Q_s = 0.59$) and 5387 pairs ($Q_s = 0.4$) local associations were shared, respectively. On the other hand, significant local associations are not always shared in large number between different sites even within one person. Only 2 pairs ($Q_s = 0.18$) were shared between the feces and tongue sites for 'F4' while 22 pairs ($Q_s = 0.48$) for 'M3'. However, owing to the small number of total shared OTUs between the two sites, the Sorensen indices were not low. In addition, compared with 'F4', 'M3' generally has a more similar microbiota across body sites, since, between the feces and palm sites for 'M3', 245 pairs were shared ($Q_s = 0.75$) and between tongue and palm sites 605 pairs were shared ($Q_s = 0.67$), while, between the feces and palm sites for 'F4', only 34 pairs were shared ($Q_s = 0.37$) and between tongue and palm sites only 73 pairs were shared ($Q_s = 0.4$).

Among the significant local associations we found, many of them are time-delayed local associations. For example, in the 'F4' feces sample, the global profiles *Coprococcus* and *Escherichia* are not significantly correlated by PCC ($r = -0.1708$, $P = 0.0521$) while significantly by eLSA (LS $= -0.3179$, $P = 0.0002$). The association is significantly negative for 126 consecutive time points, where *Coprococcus* leads *Escherichia* by 3 days. Hinted by this, shifting the *Coprococcus* profile by 3 days backward, we see their global profile are significantly negatively correlated ($r = -0.3314$, $P = 0.0001$) (Supplementary Figure S2). As another example, in the 'F4' feces sample, *Eubacterium* and *Oscillospira* are not significantly correlated by PCC ($r = 0.1313$, $P = 0.1364$), however, significantly associated for 122 consecutive time points by eLSA (LS $= 0.3862$, $P = 0.0001$), where the former leads the later by 2 days in the co-occurrence. Hinted by this, shifting the *Eubacterium* profile by 2 days backward, we see they actually are also globally significantly correlated ($r = 0.3525$, $P = 0.0001$) (Supplementary Figure S2).

4 DISCUSSION

In this article, we provide theoretical formulas to approximate the statistical significance of local similarity analysis for time series data, which make it possible to evaluate the *P*-values of comparisons of time series data for a large number of factors such as genes in gene expression analysis or OTUs in metagenomic studies, originally impractical to carry out using the permutation-based approach. The theoretical approximation is mathematically sound with specified assumptions of data distributions verifiable before the analysis. The permutation test, however, heavily depends on data-specific empirical distributions and can be biased by the numerical properties of specific data as well as its intrinsic variability.

In addition, if we are interested in the tail distribution as in most applications, the permutation and theoretical methods are mostly in agreement with each other in predictions given the same type-I error threshold. We have results from setting threshold to lower values (0.01, 0.005, 0.001, etc.) showing high overall agreement rate (data not shown). Therefore, from the practical point of view, we can substitute permutations with the theoretical method in such applications. Moreover, from the simulations and our real analysis, P_{theo} is more conservative than P_{perm}—a property particularly useful in biological applications prone to substantial number of false positives, such as the microarray analysis (Pawitan *et al.*, 2005).

The most important reason for us to embrace the theoretical method is computational efficiency. As shown in Xia *et al.* (2011b), for a given type-I error, α, the time complexity of computing P_{prem}, is $O(DMN/\alpha)$, where D is the delay limit, N is the sample number and M the replicate number. With P_{theo}, we may compute and store (LS score, *P*-value) pairs into a hash table, before any pairwise comparison. Then, for each comparison, it only costs constant time $O(1)$ to read out P_{theo} and is independent of D, M, N and α, a strongly desired feature for large scale analysis. The superiority of efficiency is evident from Supplementary Figure S3, in which, the averaged time cost of analysing four sets of 40 factors, 114 time points series was compared between theoretical approach and 1000 times permutation. The per-pair time cost is about 40 seconds for P_{perm} while negligible for P_{theo} and is independent of sample size, which is a big saver of computing resource, energy and research time.

5 CONCLUSIONS

The recent advent of high-throughput technologies made possible large scale time-resolved omics studies (proteomics, transcriptomics, metagenomics), tracking hundreds, thousands or even tens of thousands of molecules simultaneously. Time-series generated from these studies provide an invaluable opportunity to investigate the varying dynamics of biological systems. However, to make full use of huge datasets, accurate and efficient statistical and computational methods are urgently needed at all levels of analysis, from accurate estimation of abundance and expression levels, to pairwise association and network analysis.

The theoretical statistical significance approximation we proposed in this work can serve as an efficient alternative for calculating *P*-values in local similarity analysis. Its time cost is always constant, which reduces the computational burden in a large scale pairwise analysis. For example, in metagenomics, after short read assignment and abundance estimation (Xia *et al.*, 2011a; He and Xia, 2007), profiles of thousands of microbial OTUs are available. Before this work, pairwise local association analysis with this number of factors was hardly tractable using permutation procedures, if not impossible. Parallel computation and hardware acceleration or some pre-clustering and filtering approaches were required, increasing the difficulty of analysis. With the new method, researchers can quickly compute the statistical significance for all OTU pairs on desktop computers, allowing on-the-fly network mining and analysis.

After analysing the MPH dataset with the new method, we found body-site-specific human microbiota core OTUs are highly coordinated. There exist robust site-specific associations across individuals. We implemented the new method in the eLSA package (Xia *et al.*, 2011b), which now provides faster pipelines

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023

for local similarity analysis. The tool is freely available from eLSA's website: http://meta.usc.edu/softs/lsa.

ACKNOWLEDGEMENT

We sincerely thank Professor Ken Alexander from the Mathematics Department at USC for pointing us to the reference Feller (1951).

Funding: This research is partially supported by US NSF DMS-1043075 and OCE 1136818, and National Natural Science Foundation of China (60928007 and 60805010).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Androulakis,I.P. *et al.* (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, **9**, 205–228.

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

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Beman,J.M. *et al.* (2011) Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal california. *ISME J.*, **5**, 1077–1085.

Caporaso,J.G. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.

Chaffron,S. *et al.* (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.*, **20**, 947–959.

Daudin,J.J. *et al.* (2003) Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stoch. Proc. Appl.*, **107**, 1–28.

Durno,W.E. *et al.* (2013) Expanding the boundaries of local similarity analysis. *BMC Genomics*, **14** (**Suppl. 1**), S3.

Etienne,M.P. and Vallois,P. (2004) Approximation of the distribution of the supremum of a centered random walk application to the local score. *Methodol. Comput. Appl.*, **6**, 255–275.

Feller,W. (1951) The asymptotic distribution of the range of sums of independent random variables. *Ann. Math. Stat.*, **22**, 427–432.

Gilbert,J.A. *et al.* (2011) Defining seasonal marine microbial community dynamics. *ISME J.*, **6**, 298–308.

He,F. and Zeng,A.P. (2006) In search of functional association from time-series microarray data based on the change trend and level of gene expression. *BMC Bioinformatics*, **7**, 69.

He,P. and Xia,L. (2007) Oligonucleotide profiling for discriminating bacteria in bacterial communities. *Comb. Chem. High T. Scr.*, **10**, 247–255.

Ji,L. and Tan,K.L. (2005) Identifying time-lagged gene clusters using gene expression data. *Bioinformatics*, **21**, 509–516.

Karlin,S. *et al.* (1990) Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.*, **18**, 571–581.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90**, 5873–5877.

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

Pawitan,Y. *et al.* (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

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.

Quinn,G.P. and Keough,M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK; New York.

Ruan,Q. *et al.* (2006) Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, **22**, 2532–2538.

Shade,A. *et al.* (2010) Differential bacterial dynamics promote emergent community robustness to lake mixing: an epilimnion to hypolimnion transplant experiment. *Environ. Microbiol.*, **12**, 455–466.

Sorensen,T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Biol. Krifter Bd.*, **4**, 1–34.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Steele,J.A. *et al.* (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J., **5**, 1414–1425.

Xia,L.C. *et al.* (2011a) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, **6**, e27992.

Xia,L.C. *et al.* (2011b) Extended local similarity analysis (elsa) of microbial community and other time series data with replicates. *BMC Syst. Biol.*, **5** (**Suppl. 2**), S15.

Downloaded from https://academic.oup.com/bioinformatics/article/29/2/230/202138 by guest on 16 November 2023