EFFICIENT STATISTICAL SIGNIFICANCE APPROXIMATION FOR LOCAL ASSOCIATION ANALYSIS OF HIGH-THROUGHPUT TIME SERIES DATA

by

Li Charlie Xia

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(STATISTICS)

August 2012

Dedication

To my wife Ge, my parents Songjuan and Xianhui, and to my professors and friends.

Acknowledgements

I am most grateful to my advisor, Prof. Fengzhu Sun. Without his insightful vision and continuous support, I would not have come this far. I am also deeply impressed and influenced by his rigorous academic attitude, for instance, careful examination of every line of my manuscripts, which I will carry on these merits I learned from him to my future research and life.

I also would like to thank Prof. Allan Schumitzky and Prof. Jianfeng Zhang, who served as my MS committee. With their mathematical expertise, they have been valuable resource to resort to and particular helpful during my research. The questions and comments raised by them directed me to further improve my works.

I want to express my thanks to Profs. Sun and Chen's current and previous group members: Lin, Xiting, Joyce, Jing, Quan, Wangshu, Xuemei, Yang-ho, Kjong, Sungjie, Xiaolin, Tade and others, and Prof. Fuhrman group, Joshua, Jacob, Cheryl, Rohan and others I cannot enumerate. Thank you for your company through my PhD journey. Finally, I want to express my thanks to my family, other USC faculty members and friends for sharing with me a period of wonderful time in Los Angeles. Thank you all!

Table of Contents

List of Tables

List of Figures

Abstract

Local association analysis, such as local similarity analysis and local shape analysis, of biological time series data helps elucidate the varying dynamics of biological systems. However, their applications to large scale high-throughput data are limited by slow permutation procedures for statistical significance evaluation. We developed a theoretical approach to approximate the statistical significance of local similarity and local shape analysis based on the approximate tail distribution of the maximum partial sum of independent identically distributed (i.i.d) and Markovian 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 to those from permutations. The new approach enables efficient calculation of statistical significance for pairwise local association analysis, making possible all-to-all association studies otherwise prohibitive. As a demonstration, local association analysis of human microbiome time series shows that core OTUs are highly synergetic and some of the associations are body-site specific across samples. The new approach is implemented in our eLSA package, which now provides pipelines for faster local similarity and shape analysis of time series data. The tool is freely available from eLSA's website: http://meta.usc.edu/softs/lsa.

Chapter 1

Introduction

1.1 Local association analysis approaches

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, see reviews [1, 3]. 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 relationships of the gene expression profiles may fail to detect these relationships.

Several methods have been developed to address this problem. Borrowing the idea from local alignment for molecular sequences, [18] 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. [20] used a similar approach to study local associations of microbial organisms in the ocean over a four year period and this approach has been used in several other recent ecological studies [4, 6, 10, 21, 23]. [25] 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.

Several investigators have extended the basic local similarity approach for the gene expression profiles to local shape analysis [2, 12, 14], where in [14], the time delay is fixed and in [12], the time delay is not pre-defined but is estimated from the data. In local shape analysis, expression profiles of n consecutive time points are changed to a $n-1$ time point series corresponding to decrease, no change, and increase in expression levels. The Smith-Waterman algorithm is then used to locate subintervals I and J in the pair of obtained series, respectively, so that the product is maximized. The statistical significance is evaluated by permutation procedures similar to that of local similarity analysis. However, this approach is problematic since the shape data are not independent even under the assumption that the real observed data are independent.

1.2 Current limitations and our developments

One of the major limitations of the local similarity/shape 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 is prohibitive. While in practice false discovery rate (q-value) is used to mitigate the multiple comparison problem, still, fast and efficient theoretical approximation for the statistical significance of the LS score is urgently needed to accurately estimate the p-values.

In this thesis work, we provide a theoretical approximation for the distribution of local similarity analysis LS scores as well as that based on local shape analysis. In the "Methods" chapter, we provide the theoretical bases for deriving the approximate tail probability that the LS score is above a threshold. In the "Results" chapter, 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 thesis concludes with some discussion on further applications and future research directions.

Chapter 2

Methods

2.1 Previous works

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 [18, 20, 25]. 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 apart, i.e. $|i - j| \leq D$. We denote the LS score with time delay at most D 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 [18, 20]. 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 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.

The local shape analysis [2, 12, 14] is similar to local similarity analysis given above. The only difference is to change the vectors of expression levels to a vector of {-1, 0, 1} corresponding to decrease, no significant change, or increase of the expression levels at consecutive time points. Under this situation, the permutation approach for the obtained vectors is even more problematic because the adjacent terms are highly correlated and the permutation approach destroys such dependency resulting in incorrect estimation of p-values. We will also provide new theoretical approaches to calculate the p-value corresponding to local shape analysis.

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

In order to derive theoretical formulas to calculate the p-value related to the local similarity score, we resort to classical theoretical studies on the range of partial sums for independent identically distributed (i.i.d) random variables with zero mean [9] and the extensions to Markovian random variables [7]. 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) which forms a milestone in the field of computational biology [15]. On the other hand, the theoretical results on the approximate distributions when the mean is zero is not widely 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. [9] 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 [9]: $E(R_n/\sigma) = 2\sqrt{2n/\pi}$, $\mathrm{Var}(R_n/\sigma) = 4n(\log(2) - 2/\pi)$.

Using the theory of Bachelier-Wiener processes, [9] approximated the density function of R_n/σ by (equations 3.7 and 3.8 in that paper) $\delta(n; r)$,

$$\delta(n; r) = \sqrt{\frac{2}{\pi}} r^{-1} L'(r/(2\sqrt{n})), \tag{2.1}$$

where,

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

Thus,

$$\begin{aligned} & P(R_n/(\sigma\sqrt{n}) \geq x) \\ = & \int_{\sqrt{n}x}^{\infty} \sqrt{\frac{2}{\pi}} r^{-1} L'\left(\frac{r}{2\sqrt{n}}\right) dr \\ = & 1 - 8\sum_{k=0}^{\infty} \left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right) \exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right) \end{aligned} \tag{2.2}$$

Next we give an upper bound for the tail in equation 2. This upper bound can be used to determine when we stop the summation in equation 2 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,

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

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

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

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

$$1 - 8\sum_{k=0}^{K-1}\left(\frac{1}{x^2} + \frac{1}{(2k+1)^2\pi^2}\right)\exp\left(-\frac{(2k+1)^2\pi^2}{2x^2}\right). \tag{2.3}$$

[7] studied the distribution of the maximum partial sum of an aperiodic Markov chain taken values on a finite subset of the real line, i.e $H_n = \max_{0\leq i<j\leq n}(S_j - S_i)$. 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 [7] that

$$\lim_{n\to\infty} P\left(\frac{H_n}{\sigma\sqrt{n}} \leq x\right) = \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right). \tag{2.4}$$

[8] provided an upper bound of $C\sqrt{\log(n)/n}$ for the approximation in equation 2.4.

Similarly, we can define $L_n = -\min_{0\leq i<j\leq n}(S_j - S_i)$. 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 $\{H_n > x\}\bigcap\{L_n > x\}$ will be small and

$$\begin{aligned}
& P(R_n/(\sigma\sqrt{n}) \geq x) \\
= \;& P(\max(H_n, L_n)/(\sigma\sqrt{n}) \geq x) \\
= \;& P(\{H_n/(\sigma\sqrt{n}) \geq x\}\bigcup\{L_n/(\sigma\sqrt{n}) \geq x\}) \\
\approx \;& P(H_n/(\sigma\sqrt{n}) \geq x) + P(L_n/(\sigma\sqrt{n}) \geq x) \\
\approx \;& 2P(H_n/(\sigma\sqrt{n}) \geq x) \\
\approx \;& 2\left(1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right)\right), \quad x \geq 0.
\end{aligned} \tag{2.5}$$

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

2.3 Statistical significance for local similarity scores

We next use the theory outlined in section 2.2 to approximate the statistical significance in local similarity and local shape analyses. For time series data of two factors $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_n$, 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-value $= P(LS(D) \geq s_D)$, where $LS(D) = \max_{i,j,l;|i-j|\leq D} |\sum_{k=0}^{l-1} X_{i+k}Y_{j+k}|$. First consider the case that $D = 0$. Let $Z_i = X_iY_i, i = 1, 2, \cdots, n$. Assuming that both X_i and Y_i are independent standard normally distributed, we have $E(Z_i) = 0$ and $E(Z_i^2) = 1$. Therefore, we can directly use the theory developed above in equations 2.3 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 $(d = 0, \pm 1, \pm 2, \cdots, \pm D)$

$$\begin{array}{ccccccc} x_1 & x_2 & x_3 & \cdots\cdots\cdots & x_{n-2} & x_{n-1} & x_n \\ y_{1+d} & y_{2+d} & y_{3+d} & \cdots\cdots\cdots & x_{n-2+d} & x_{n-1+d} & x_{n+d} \end{array}$$

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 2.3. Note $LS(D) =$

$\max_{d=-D}^{D} S_n^{(d)}$. In order 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

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

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

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

From Equation 2.6, we can obtain the approximate density function of $R_n^{(D)}/(\sigma\sqrt{n})$ by

$$\begin{aligned} f_D(x) = & \frac{(2D+1)}{x^3} 8^{2D+1} \left(\sum_{k=1}^{\infty} \left(\frac{1}{x^2} + \frac{1}{(2k-1)^2\pi^2} \right) \exp\left(-\frac{(2k-1)^2\pi^2}{2x^2} \right) \right)^{2D} \\ & \times \sum_{k=1}^{\infty} \left(\frac{(2k-1)^2\pi^2}{x^2} - 1 \right) \exp\left(-\frac{(2k-1)^2\pi^2}{2x^2} \right). \end{aligned} \tag{2.6}$$

Finally, we can model time series data with replicate data as well as with large fraction of zero values based on equation 2.6, using additional scaling in σ, and the detail of the method is given in the Supplementary Methods.

2.4 Statistical significance for local shape analysis

Next, we use the theory developed in section 2.2 to approximate the p-value corresponding to local shape (also LS) score. Note that in local shape analysis, we change the n-dimensional vector X to a $n-1$ dimensional vector $(d_i^X, i = 1, 2, \cdots, n-1)$, where $d_i^X = \text{sign}(X_{i+1} - X_i)$, $\text{sign}(x) = 1$ when $x > 0$ and $\text{sign}(x) = -1$ when $x < 0$. Note that even if the components of X are independent, the random variables $d_i^X, i = 1, 2, \cdots, n-1$ are not independent because both d_i^X and d_{i+1}^X depend on X_{i+1}. Assuming that $X_1, X_2, \cdots, X_n$ are i.i.d continuous random variables such that the probability of taking a fixed value to be 0, we have $P[(d_i^X, d_{i+1}^X) = (1,1)] = P[(d_i^X, d_{i+1}^X) = (-1,-1)] = 1/6$ and $P[(d_i^X, d_{i+1}^X) = (1,-1)] = P[(d_i^X, d_{i+1}^X) = (-1,1)] = 1/3$.

Note that $P(d_i^X = 1) = P(d_i^X = -1) = 1/2$ if the X's are exchangible. Therefore $d_i^X,\ i = 1, 2, \cdots, n-1$ are not independent. Actually $d_1^X, d_2^X, \cdots, d_{n-1}^X$ do not even form a Markov chain. However, let us pretend that they form a Markov chain $\overline{d}_i^X$ with transition matrix

$$T = \begin{array}{c|cc} & 1 & -1 \\ \hline 1 & 1/3 & 2/3 \\ -1 & 2/3 & 1/3 \end{array} .$$

Then it can be shown by spectral expansion that

$$T^k = \frac{1}{2} \begin{pmatrix} 1 + (-1)^k/3^k & 1 - (-1)^k/3^k \\ 1 - (-1)^k/3^k & 1 + (-1)^k/3^k \end{pmatrix}.$$

For $k > 1$, we have $P(\overline{d}_1^X \overline{d}_{k+1}^X = 1) = (1 + (-1)^k/3^k)/2$ and $P(\overline{d}_1^X \overline{d}_{k+1}^X = -1) = (1 - (-1)^k/3^k)/2$. Thus, $E(\overline{d}_1^X \overline{d}_{k+1}^X) = (-1)^k/3^k$. In local shape analysis, we compare $d_1^X, d_2^X, \cdots, d_{n-1}^X$ with $d_1^Y, d_2^Y, \cdots, d_{n-1}^Y$. The score for the local shape analysis is the range of $\sum_i d_i^X d_i^Y$ when $D = 0$. Using the results of [7], we have $\sigma_{XY}^2 = E((\overline{d}_1^X)^2)E((\overline{d}_1^Y)^2) + 2\sum_{k=1}^{\infty} E(\overline{d}_1^X \overline{d}_{k+1}^X)E(\overline{d}_1^Y \overline{d}_{k+1}^Y) = 1 + 2\sum_{k=1}^{\infty} 1/3^{2k} = 1 + 1/4 = 1.25$. Thus, for the local shape score,

$$P(LS(D) \geq s_D) = P\left(\frac{LS(D)}{\sigma_{XY}\sqrt{n}} \geq \frac{s_D}{\sigma_{XY}\sqrt{n}}\right) = \mathcal{L}\left(\frac{s_D}{\sqrt{1.25 \times n}}\right),$$

where the function $\mathcal{L}$ is defined in equation 2.6.

2.5 Simulation studies and application to real datasets

In deriving the approximate p-values for local similarity and local shape analysis in sections 2.3 and 2.4, 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 2.6 using simulations for both local similarity and shape 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), $i = 1, 2, \cdots, n$ and X_i and Y_i are independent. Secondly, the dynamic programming algorithm is then 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 2.6 for local similarity scores ($\sigma = 1$). We did similar simulation studies for local

shape scores, using the transformed series of $\{-1, 0, 1\}$ from the original standard normal series ($\sigma = \sqrt{1.25}$).

We then apply our method to analyze three real datasets. The first one is a microarray yeast gene expression dataset, synchronized by cdc-15 gene, from [22] (referred to as 'CDC'). The second one is an ARISA molecular finger printing microbial ecology dataset from San Pedro Ocean Time Series in [23] (referred to as 'SPOT') . The third one is a 16S RNA tag-sequencing dataset from Moving Pictures of Human microbiota sampling of human symbiotic microbial communities from [5] (referred to as 'MPH'). We applied both local similarity and shape analysis to re-analyze the first two datasets and compared the theoretical and permutation approaches. We are the first to analyze the third dataset for local similarity and shape analyses. For the local similarity analysis, datasets are normalized by the procedures described in Supplementary Methods (referred to as 'original'). For the local shape analysis, datasets are converted to local trend series of $\{-1, 0, 1\}$ and analyzed without normalization (referred to as 'trendBy0').

Chapter 3

Results

3.1 Simulations

The approximate p-values for the local similarity and local shape analysis given in sections 2.3 and 2.4 are 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 3.1 gives the theoretical tail probability based on equation 2.6 (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 $x \geq 3$ corresponding to the theoretical p-value less than 0.011. 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 $D = 0$. Similar results are observed for local shape analysis (Table 3.2) when n is as small as 20. When $D > 0$, the theoretical approximation is close to the simulated tail probability when $n \geq 20$ and $n \geq 30$ for local similarity and shape analysis, respectively

(see Tables SB.1-B.8 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 the $R_n^{(D)}/(\sigma\sqrt{n})$ is stochastically increasing and that as n increases the theoretical approximation become closer to the simulated distribution of $LS(D)/(\sigma\sqrt{n})$. We also tested if the $P(LS(D)/(\sigma\sqrt{n}) \geq x) = 1 - (1 - P(R_n^{(0)}/(\sigma\sqrt{n}) \geq x))^{2D+1}$ is generally true using the simulated tail probabilities and it can be clearly seen from Tables SB.1-B.8 that this relationship is indeed reasonable indicating that $S_n^{(d)},\ d = 0, \pm 1, \cdots \pm D$ can indeed be considered as independent.

In equation 2.6, 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 values of $LS(D)/(\sigma\sqrt{n})$ at $n = 200$ and $D = 0, 1, 2, 3$ as in Figures 3.1-3.2 for local similarity and local shape scores, respectively. Several observations can be made from the figure. First, the values of $LS(D)/(\sigma\sqrt{n})$ increases as a function of D as expected. Second, the approximate theoretical 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 model 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 to that of permutation (P_{perm}) given the same data, see Figures SB.1-B.2. For local

x	Theory	Number of time points n 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

Table 3.1: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation 2.3 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. $D = 0$.

similarity case, 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. Similar trend is observed for local shape analysis. This indicates an increasing rate of agreement between the theoretical and permutation p-values, representing their reasonable approximation to the null distribution despite of inherent variance associated with the permutation procedures. The same is true with $D > 0$ for both analysis 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 to that from permutation starting from $n = 20$.

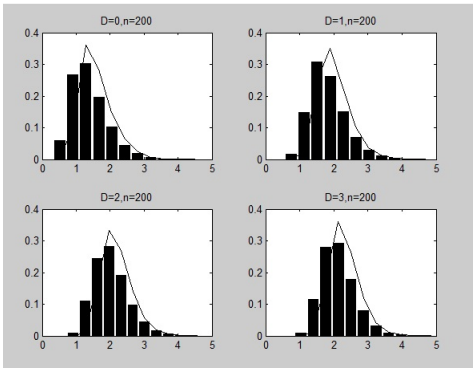

Figure 3.1: The histogram of local similarity scores $LS(D)/\sqrt{n}$ for $n = 200$ and $D = 0, 1, 2, 3$ together with the theoretical approximate density function given in Equation 2.6.

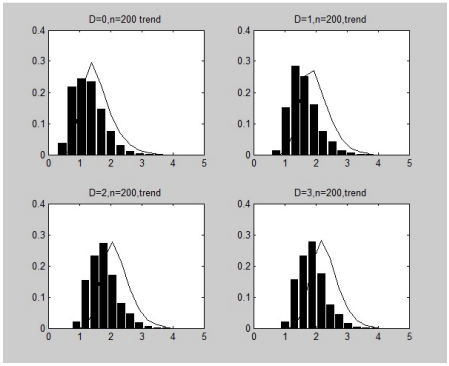

Figure 3.2: The histogram of local shape scores $LS(D)/\sqrt{1.25 \times n}$ for $n = 200$ and $D = 0, 1, 2, 3$ together with the theoretical approximate density function given in Equation 2.6.

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 the p-values by our theoretical approximation denoted by P_{theo} with the p-values by the permutation approach denoted as P_{perm}. The results are given in Figure 3.3. It can be seen from the figure that P_{theo} is highly positively associated 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}. For a specific example, Table SB.9 in the supplementary materials compares the gene pairs declared as significant by either P_{theo} or P_{perm} for the type-I error threshold 0.05. 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 ('original'), 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.

For local shape analysis ('trendBy0'), the value of P_{theo} is more closer to the value of P_{perm}. With D=0 and type-I error 0.05, we have 53 (18%) out of 300 found significant while 241 (80%) non-significant by both approaches, and in total 296 (98%) are in agreement. Among the gene pairs with $P_{perm} > 0.05$, none of them are significant using P_{theo}.

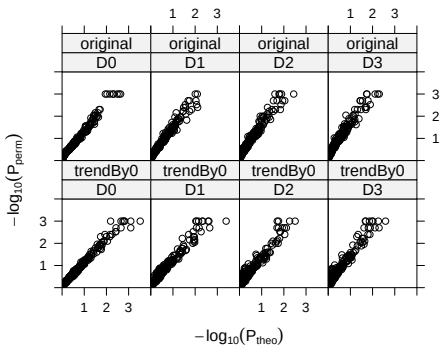

Figure 3.3: P_{theo} and P_{perm} comparison for all-to-all pairwise local similarity ('original') and shape ('trendBy0') analysis of 25 selected factors from CDC dataset. Columns D0 to D3 are for $D = 0, 1, 2, 3$, respectively.

Among the gene pairs declared as significant by P_{perm}, about 53/59 (90%) are declared as significant by P_{theo}. Similarly, with D=1,2,3, there are 291 (97%), 281 (94%) and 278 (93%) p-value pairs in agreement by both P_{perm} and P_{theo}, respectively. In fact, 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, 2.5GHz, 12010MB RAM" computing node, while, using the theoretical approach, the analysis finishes in 10 hours on the same node.

3.3 The SPOT dataset

The SPOT dataset consists of ten-year monthly (114 time points) sampled operational taxonomic unit (OTU) abundance data. A major numerical difference between the CDC and the SPOT dataset is that the later has a large number of zeros, where the OTU's abundance drops below the measurement limit. We adjusted our model for this fact as described in Supplementary Methods. As in the above section, 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". In the selected set of OTUs, we observe the fraction of zeros ranging from 0% to 46% with median at 12%, while there is almost none zeros in CDC. We summarize p-value comparison for local similarity and shape analysis in Table SB.10 and Figure SB.3 in the supplementary materials.

For local similarity analysis ('original'), 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_{theo} but non-significant by P_{perm}. The results are similar with D=1,2,3: 733 (94%), 723 (93%) and 727 (93%) in concordance, respectively. There are about 6-7% associations significant by P_{perm} but non-significant by P_{theo}, showing that P_{theo} is more conservative. For local shape analysis ('trendBy0'), the result show even better concordance. With $D = 0$ and type-I error 0.05, we have 578 (74%) and 182 (23%) out of 780 found significant and non-significant, respectively, by both methods. In total 760 (97%) are in agreement. Among the 20 (3%) OTU pairs with discordant significance by P_{theo} and P_{perm}, all of them are significant by P_{perm} but non-significant by P_{theo}. The results are similar with D=1,2,3, with 754 (97%), 757 (97%) and 749 (96%) in concordance, respectively, and about 3-4% incidences significant by P_{perm} but non-significant by P_{theo}. The concordance for 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, 372 samples from a male ('M3') [5]. The genus level OTU abundance is used for our analysis. There are 335, 1295, 373 unique OTUs from feces, palm and tongue sites of 'F4' and 'M3', respectively. With $D = 3$, we analyzed the MPH dataset with local similarity and shape analyses. Because of the intra-person variability (both time and site) of the human microbiota, one important step in analyzing 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 [5], we consider core OTUs as those showing in at least 60% of samples from the same body site of a specific 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'.

Subsequent analysis show these symbiotic core microbes are highly synergetic. We use local shape analysis as our main approach and report significant local associations with p-value<=0.05, Q-value<=0.05 and Aligned Length >= 80% time points [25]. We found 159 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 divided by number of OTUs minus one in an association subnetwork) is 16%. The same rates are 26% and 25% for 'F4' tongue and palm samples, whereas 44%, 56% and 34% for 'M3' feces, tongue and palm samples. These percentages translate into a picture of high connectivity between these OTUs, supporting their role as crucial players in the corresponding microbiota.

However, do these site-specific significant local associations persist across individuals? In fact, we found some of the body-site specific dynamics are shared across human subjects. We take the intersection (shared) of all significant association pairs between two samples and calculate the shared percentage dividing by the union between the two. We found 'F4' and 'M3' share 83 (9%) in their feces microbiota. These are mostly from the *Ruminococcaceae, Lachnospiraceae, Clostridiaceae, Clostridiaceae Family XI Incertae Sedis, Prevotellaceae, Coriobacteriaceae* families. In tongue and palm sites, the fractions are 17% and 21%, respectively. In contrast, significant associations are generally not shared between different body sites even within one person. For example, only 1 ($\sim 0\%$) pair shows significance in both feces and tongue sites of 'F4'. Similar low fractions were observed for other sites.

On the other hand, using local similarity analysis, we found some time-delayed local associations. For example, in '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* 3 time points in the co-occurrence. Hinted by this, shifting the *Coprococcus* profile 3 time units backward, we see their global profile are significantly negatively correlated (r=-0.3314, P=0.0001), see Figure SB.4. As another example in 'F4' feces sample, *Eubacterium* and *Oscillospira* is 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 2 time points in the co-occurrence. Hinted by this, shifting the *Eubacterium* profile 2 time units backward, we see they actually are also globally significantly correlated (r=0.3525, P=0.0001), see Figure SB.4.

3.5 Discussion

In this paper, we provide theoretical formulas to approximate the statistical significance of local similarity and local shape analyses for time series data. The theoretical approximations make it possible to evaluate the statistical significance of comparisons of time series data for a large number of factors such as genes in gene expression analysis or OTUs in metagenomic studies, which is impossible to carry out using the original permutation based approach. The theoretical approximation is more mathematically sound with specified assumptions of data distribution verifiable before applying 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 particular, in the local shape analysis, permuting the sequence independently may be misleading because the trend sequence is generated with dependency.

In addition, if we are interested in the tail distribution (in most applications), the two 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. 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 P_{perm}

– a property particularly useful in biological applications prone to substantial number of false positives, such as the microarray analysis [17].

The most important reason for us to embrace the theoretical method is computational efficiency. As shown in [25], for a given type-I error, α, the time complexity of P_{perm} 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 them into a hash table, before any pairwise comparison. Then, for each p-value calculation, it only costs constant time $O(1)$ to read out P_{theo} and is independent of D, M, N and Q, a strongly desired feature for large scale analysis. The superiority of efficiency is evident from Figure SB.5, in which, the time cost of analyzing 40 factors of 113 and 114 time points are compared. The per-pair time cost is about 40 seconds for P_{perm} while negligible for P_{theo} and independent of sample size, which is a big saver of computing resource, energy, and research time.

		The number of time points n						
x	Theory	10	20	30	40	60	80	100
2	0.1815	0.0491	0.1242	0.0921	0.0986	0.1092	0.1439	0.1308
2.2	0.1111	0.0491	0.0677	0.0613	0.0680	0.0599	0.0832	0.0801
2.4	0.0656	0.0132	0.0367	0.0353	0.0430	0.0420	0.0492	0.0487
2.6	0.0373	0.0053	0.0168	0.0211	0.0158	0.0207	0.0266	0.0207
2.8	0.0204	0.0053	0.0094	0.0057	0.0106	0.0098	0.0137	0.0125
3.0	0.0108	0.0000	0.0039	0.0027	0.0034	0.0057	0.0069	0.0080
3.2	0.0055	0.0000	0.0016	0.0013	0.0021	0.0026	0.0032	0.0044
3.4	0.0027	0.0000	0.0002	0.0007	0.0009	0.0009	0.0008	0.0020
3.6	0.0013	0.0000	0.0002	0.0000	0.0004	0.0005	0.0004	0.0009
3.8	0.0006	0.0000	0.0000	0.0000	0.0001	0.0003	0.0001	0.0004
4.0	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
4.2	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
4.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3.2: Theoretical approximation for local shape analysis p-values versus the simulated probability $P(LS(D)/\sqrt{1.25 \times n} \geq x)$. The theoretical approximate probability based on equation 2.3 with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{1.25n} \geq x$ is given in the 3rd to the 9th columns. $D = 0$.

Chapter 4

Conclusions

4.1 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 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 in 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 and shape 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 [13, 24], profiles of thousands of microbial OTUs are available. Before this work, pairwise local association analysis with this number of factors are hardly tractable using permutation procedures, if not impossible. Parallel computation and hardware acceleration or some pre-clustering and filtering approaches are 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.

Analyzing 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 persons. We implemented the new method in the eLSA package [25], now providing faster pipelines for local similarity and shape analysis. The methodological part remain true for other types of local associations that use the maximum range of partial sum of i.i.d and markovian zero-mean and finite variance random variables as their metrics. It may be further developed for application in more complex analysis, such as local shape with non-zero thresholds and local liquid association analysis.

Bibliography

[1] I. P. Androulakis, E. Yang, and R. R. Almon. Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu Rev Biomed Eng*, 9:205–28, 2007.

[2] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–77, 2005.

[3] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–503, 2004.

[4] J. M. Beman, J. A. Steele, and J. A. Fuhrman. Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal california. *ISME J*, 5(7):1077–85, 2011.

[5] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, et al. Moving pictures of the human microbiome. *Genome Biol*, 12(5):R50, 2011.

[6] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res*, 20(7):947–59, 2010.

[7] J. J. Daudin, M. P. Etienne, and P. Vallois. Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stoch Proc Appl*, 107(1):1–28, 2003.

[8] M. P. Etienne and P. Vallois. Approximation of the distribution of the supremum of a centered random walk. application to the local score. *Methodol Comput Appl*, 6(3):255–275, 2004.

[9] W. Feller. The asymptotic distribution of the range of sums of independent random variables. *Ann Math Stat*, 22(3):427–432, 1951.

[10] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbruck, J. Reeder, B. Temperton, S. Huse, A. C. McHardy, R. Knight, I. Joint, et al. Defining seasonal marine microbial community dynamics. *ISME J*, 6(2):298–308, 2011.

[11] M. S. Gilthorpe, M. Frydenberg, Y. Cheng, and V. Baelum. Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in medicine*, 28(28):3539–53, 2009.

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

[13] P. He and L. Xia. Oligonucleotide profiling for discriminating bacteria in bacterial communities. *Comb Chem High T Scr*, 10(4):247–255, 2007.

[14] L. Ji and K. L. Tan. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics*, 21(4):509–16, 2005.

[15] S. Karlin, A. Dembo, and T. Kawabata. Statistical composition of high-scoring segments from molecular sequences. *Ann Stat*, 18(2):571–581, 1990.

[16] K. C. Li. Genome-wide coexpression dynamics: theory and application. *Proc Natl Acad Sci U S A*, 99(26):16875–80, 2002.

[17] Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–24, 2005.

[18] J. Qian, M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J Mol Biol*, 314(5):1053–66, 2001.

[19] G. P. Quinn and M. J. Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge, UK ; New York, 2002.

[20] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, 22(20):2532–8, 2006.

[21] A. Shade, C. Y. Chiu, and K. D. McMahon. Differential bacterial dynamics promote emergent community robustness to lake mixing: an epilimnion to hypolimnion transplant experiment. *Environ Microbiol*, 12(2):455–66, 2010.

[22] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, 1998.

[23] J. A. Steele, P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C. E. Chow, R. Sachdeva, A. C. Jones, M. S. Schwalbach, et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J*, 5(9):1414–25, 2011.

[24] L. Xia, J. Cram, T. Chen, J. Fuhrman, and F. Sun. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS ONE*, 6(12):e27992, 2011.

[25] L. Xia, J. Steele, J. Cram, Z. Cardon, S. Simmons, J. Vallino, J. Fuhrman, and F. Sun. Extended local similarity analysis (elsa) of microbial community and other time series data with replicates. *BMC Syst Biol*, 5(Suppl 2):S15, 2011.

Appendix A

Supplementary Methods

A.1 Dealing with replicates.

To reduce the effect of biological and/or technical variation on the LSA results, replicate experiments are frequently carried out [19]. An extended LSA (eLSA) [25] approach was developed for time series data with replicates. First, for each sample the replicate data at each time point was 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 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 $U_i = F(X_i^{(1)}, \cdots, X_i^{(M)}) = \overline{X}_i = \sum_{m=1}^{M} X_i^{(m)}/M$ 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 normalized to be 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

$$P(LS(D) \geq s_D) = P\left(\frac{M \times LS(D)}{\sqrt{n}} \geq M \times s_D/\sqrt{n}\right) = \mathcal{L}(M \times s_D/\sqrt{n}),$$

where the function $\mathcal{L}$ is defined in equation 2.6 in the Method section.

A.2 Data normalization

eLSA requires the input series of factors X and Y to be standard normally distributed, which 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 [16]. We use x_i to denote the original raw data of the i-th time spot of a factor X. First, we take

$$r_k = \text{rank of } x_k \text{ in } \{x_1, x_2, ..., x_n\}.$$

Then, we take

$$s_k = \Phi^{-1}\left(\frac{r_k}{n+1}\right),$$

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

$$z_i = \frac{s_i - \bar{S}}{\sigma_S}.$$

We will take $Z = z_{[1:n]}$ as the standardized normalization of X.

A.3 Dealing with multiple zeros

When normalizing the time series data, we rank the input data by their numerical value and then take the standard normal percentile which equals to their rank divided by the total number of data points plus one. In some real situations, we may encounter a number of indistinguishable zeros (beyond measurement limit) from the input, which violates our assumption that the data comes from a normal distribution.

To accommodate these indistinguishable zeros, we adjust our theoretical model using mixture ideas similar to zero-inflated models [11]. Let the proportion of zeros in the input sequences X and Y be $1-\alpha$ and $1-\beta$, respectively. In the normalization step, we bypass

these zeros, while the other non-zeros are normalized as usual. The normalized series Z_X and Z_Y thus can be modeled as i.i.d. sampled from the following mixture distributions:

$$Z_X \sim (1-\alpha)\{0\} + \alpha W_X,$$

$$Z_Y \sim (1-\beta)\{0\} + \beta W_Y,$$

where W_X and W_Y are independent and standard normal $N(0,1)$. Consequently,

$$Z_X Z_Y \sim (1-\alpha\beta)\{0\} + \alpha\beta W_X W_Y.$$

Therefore, $E\{Z_X Z_Y\} = 0$ and $Var(Z_X Z_Y) = \alpha\beta$. Fortunately, the standard theory we have developed still applies here, with a shrink in variance corresponding to the remaining non-zero portion of the product mixture distribution. Then the p-value can be calculated by:

$$P(LS(D) \geq s_D) = P\left(\frac{LS(D)}{\sqrt{\alpha\beta n}} \geq s_D/\sqrt{\alpha\beta n})\right) = \mathcal{L}\left(s_D/\sqrt{\alpha\beta n}\right),$$

where the function $\mathcal{L}$ is defined in equation 2.6 in the Method section. Similar results can be obtained for replicated series and local trend series with multiple zeros.

Appendix B

Supplementary Results

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.1815	0.0848	0.0987	0.1062	0.1122	0.1201	0.1235	0.1290	0.1294	0.1355	0.1375	0.1430	0.1379
2.2	0.1111	0.0541	0.0621	0.0645	0.0665	0.0699	0.0771	0.0767	0.0783	0.0798	0.0798	0.0861	0.0831
2.4	0.0656	0.0341	0.0367	0.0392	0.0416	0.0411	0.0435	0.0457	0.0451	0.0477	0.0483	0.0526	0.0498
2.6	0.0373	0.0223	0.0221	0.0252	0.0235	0.0232	0.0249	0.0261	0.0253	0.0261	0.0276	0.0301	0.0275
2.8	0.0204	0.0147	0.0128	0.0154	0.0131	0.0129	0.0138	0.0163	0.0129	0.0141	0.0159	0.0159	0.0152
3.0	0.0108	0.0093	0.0082	0.0088	0.0074	0.0069	0.0071	0.0090	0.0072	0.0066	0.0087	0.0083	0.0074
3.2	0.0055	0.0056	0.0051	0.0038	0.0036	0.0030	0.0035	0.0054	0.0040	0.0042	0.0043	0.0043	0.0038
3.4	0.0027	0.0033	0.0031	0.0017	0.0022	0.0009	0.0016	0.0027	0.0019	0.0018	0.0027	0.0028	0.0017
3.6	0.0013	0.0019	0.0020	0.0011	0.0014	0.0004	0.0006	0.0012	0.0011	0.0007	0.0012	0.0015	0.0008
3.8	0.0006	0.0007	0.0008	0.0006	0.0010	0.0002	0.0004	0.0009	0.0007	0.0002	0.0008	0.0008	0.0004
4.0	0.0003	0.0004	0.0005	0.0003	0.0005	0.0000	0.0003	0.0004	0.0004	0.0001	0.0002	0.0002	0.0001
4.2	0.0001	0.0002	0.0004	0.0002	0.0005	0.0000	0.0001	0.0002	0.0003	0.0000	0.0001	0.0001	0.0001
4.4	0.0000	0.0001	0.0003	0.0001	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4.6	0.0000	0.0000	0.0003	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4.8	0.0000	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	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	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.1: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation 2.3 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 14th columns. **D = 0.**

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.4516	0.1692	0.2289	0.2464	0.2698	0.2972	0.3084	0.3220	0.3354	0.3393	0.3410	0.3531	0.3491
2.2	0.2977	0.1139	0.1485	0.1636	0.1708	0.1871	0.1978	0.2043	0.2169	0.2150	0.2119	0.2300	0.2254
2.4	0.1841	0.0757	0.0976	0.0988	0.1050	0.1123	0.1176	0.1261	0.1352	0.1313	0.1292	0.1401	0.1393
2.6	0.1077	0.0513	0.0606	0.0608	0.0626	0.0650	0.0692	0.0748	0.0813	0.0776	0.0745	0.0768	0.0832
2.8	0.0601	0.0318	0.0385	0.0357	0.0379	0.0374	0.0398	0.0464	0.0460	0.0426	0.0447	0.0434	0.0470
3.0	0.0320	0.0186	0.0224	0.0187	0.0233	0.0208	0.0222	0.0276	0.0245	0.0220	0.0246	0.0249	0.0248
3.2	0.0164	0.0127	0.0124	0.0113	0.0133	0.0117	0.0115	0.0139	0.0122	0.0129	0.0128	0.0127	0.0133
3.4	0.0081	0.0078	0.0074	0.0062	0.0073	0.0058	0.0063	0.0078	0.0065	0.0066	0.0066	0.0060	0.0066
3.6	0.0038	0.0044	0.0039	0.0030	0.0045	0.0025	0.0032	0.0043	0.0039	0.0035	0.0035	0.0034	0.0034
3.8	0.0017	0.0032	0.0023	0.0015	0.0023	0.0012	0.0019	0.0023	0.0017	0.0015	0.0018	0.0015	0.0014
4.0	0.0008	0.0013	0.0013	0.0005	0.0011	0.0007	0.0010	0.0009	0.0004	0.0005	0.0007	0.0006	0.0005
4.2	0.0003	0.0010	0.0009	0.0003	0.0003	0.0002	0.0004	0.0004	0.0003	0.0005	0.0003	0.0003	0.0003
4.4	0.0001	0.0007	0.0003	0.0003	0.0001	0.0001	0.0004	0.0000	0.0002	0.0004	0.0002	0.0002	0.0002
4.6	0.0001	0.0004	0.0002	0.0003	0.0000	0.0001	0.0002	0.0000	0.0001	0.0002	0.0002	0.0001	0.0000
4.8	0.0000	0.0002	0.0002	0.0001	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000
5.0	0.0000	0.0002	0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.2: Theoretical approximation for local similarity analysis p-values versus Simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation 2.3 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 14th columns. $\mathbf{D = 1.}$

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.6326	0.2199	0.2914	0.3429	0.3704	0.4167	0.4447	0.4709	0.4792	0.4855	0.4935	0.5072	0.5178
2.2	0.4452	0.1514	0.1947	0.2316	0.2437	0.2759	0.3004	0.3132	0.3245	0.3264	0.3305	0.3429	0.3456
2.4	0.2876	0.1053	0.1228	0.1478	0.1524	0.1738	0.1884	0.1964	0.2032	0.2020	0.2043	0.2145	0.2141
2.6	0.1730	0.0713	0.0765	0.0887	0.0949	0.1036	0.1119	0.1146	0.1194	0.1203	0.1211	0.1226	0.1301
2.8	0.0981	0.0447	0.0453	0.0534	0.0558	0.0602	0.0660	0.0681	0.0657	0.0704	0.0690	0.0721	0.0732
3.0	0.0528	0.0288	0.0256	0.0303	0.0316	0.0347	0.0356	0.0385	0.0352	0.0370	0.0360	0.0421	0.0412
3.2	0.0272	0.0185	0.0139	0.0172	0.0170	0.0188	0.0199	0.0187	0.0179	0.0213	0.0185	0.0227	0.0192
3.4	0.0134	0.0121	0.0084	0.0100	0.0099	0.0097	0.0099	0.0089	0.0089	0.0092	0.0103	0.0120	0.0091
3.6	0.0063	0.0077	0.0046	0.0060	0.0051	0.0048	0.0059	0.0053	0.0044	0.0048	0.0045	0.0064	0.0053
3.8	0.0029	0.0053	0.0021	0.0037	0.0026	0.0021	0.0021	0.0031	0.0030	0.0023	0.0024	0.0035	0.0024
4.0	0.0013	0.0031	0.0011	0.0018	0.0019	0.0013	0.0007	0.0017	0.0017	0.0011	0.0014	0.0015	0.0010
4.2	0.0005	0.0020	0.0003	0.0009	0.0006	0.0006	0.0003	0.0011	0.0012	0.0003	0.0004	0.0006	0.0005
4.4	0.0002	0.0008	0.0003	0.0005	0.0005	0.0001	0.0001	0.0006	0.0006	0.0002	0.0003	0.0003	0.0001
4.6	0.0001	0.0006	0.0001	0.0004	0.0004	0.0000	0.0001	0.0003	0.0002	0.0000	0.0001	0.0001	0.0000
4.8	0.0000	0.0002	0.0000	0.0002	0.0001	0.0000	0.0001	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0001	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.3: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation 2.3 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 14th columns. $\mathbf{D = 2.}$

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.7539	0.2509	0.3331	0.4103	0.4544	0.5120	0.5402	0.5571	0.5804	0.5972	0.6128	0.6059	0.6259
2.2	0.5616	0.1731	0.2301	0.2772	0.3124	0.3533	0.3707	0.3917	0.4077	0.4109	0.4368	0.4256	0.4406
2.4	0.3779	0.1177	0.1486	0.1772	0.2000	0.2293	0.2344	0.2539	0.2613	0.2679	0.2819	0.2722	0.2805
2.6	0.2336	0.0785	0.0952	0.1122	0.1236	0.1366	0.1401	0.1578	0.1562	0.1573	0.1756	0.1678	0.1669
2.8	0.1346	0.0513	0.0583	0.0679	0.0736	0.0767	0.0813	0.0918	0.0875	0.0894	0.0977	0.0998	0.0947
3.0	0.0732	0.0320	0.0343	0.0379	0.0416	0.0443	0.0453	0.0534	0.0469	0.0481	0.0523	0.0517	0.0509
3.2	0.0379	0.0199	0.0205	0.0207	0.0210	0.0237	0.0264	0.0278	0.0246	0.0261	0.0259	0.0275	0.0271
3.4	0.0187	0.0123	0.0129	0.0116	0.0110	0.0124	0.0122	0.0142	0.0107	0.0133	0.0145	0.0145	0.0128
3.6	0.0089	0.0083	0.0073	0.0055	0.0059	0.0058	0.0061	0.0076	0.0052	0.0067	0.0068	0.0069	0.0058
3.8	0.0040	0.0047	0.0046	0.0030	0.0029	0.0029	0.0036	0.0048	0.0023	0.0031	0.0028	0.0034	0.0027
4.0	0.0018	0.0028	0.0032	0.0013	0.0015	0.0015	0.0013	0.0024	0.0005	0.0014	0.0011	0.0010	0.0011
4.2	0.0007	0.0019	0.0021	0.0004	0.0007	0.0010	0.0005	0.0008	0.0001	0.0005	0.0008	0.0003	0.0002
4.4	0.0003	0.0009	0.0015	0.0002	0.0004	0.0007	0.0002	0.0002	0.0000	0.0000	0.0003	0.0003	0.0001
4.6	0.0001	0.0006	0.0009	0.0002	0.0003	0.0003	0.0000	0.0001	0.0000	0.0000	0.0000	0.0001	0.0001
4.8	0.0000	0.0003	0.0002	0.0000	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.4: Theoretical approximation for local similarity analysis p-values versus the simulated probability $P(LS(D)/\sqrt{n} \geq x)$. The theoretical approximate probability based on equation 2.3 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 14th columns. **D = 3.**

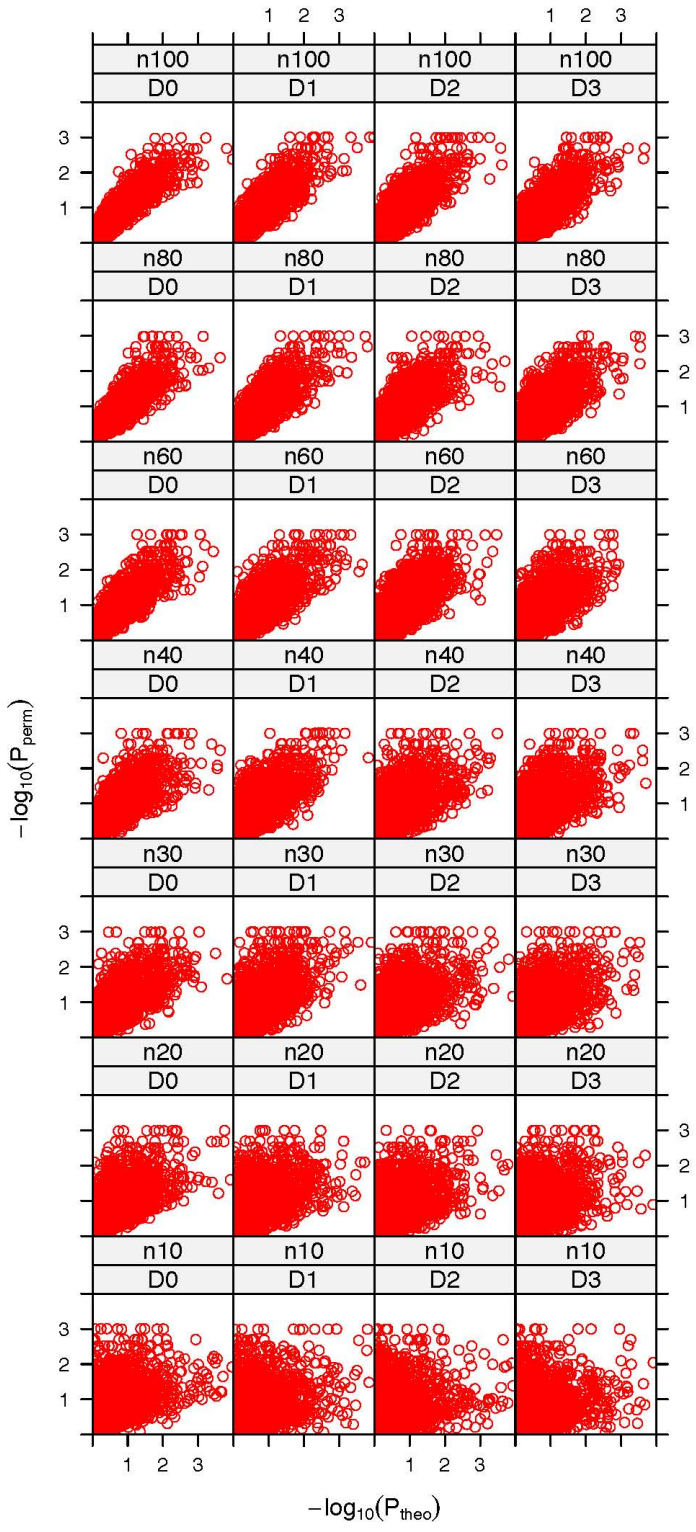

Figure B.1: Local similarity analysis P_{theo} vs P_{perm} for 10,000 pairs simulated data. Columns D0 to D3 are for $D = 0, 1, 2, 3$. Rows n10 to n100 are for $n = 10, 20, 30, 40, 60, 80, 100$.

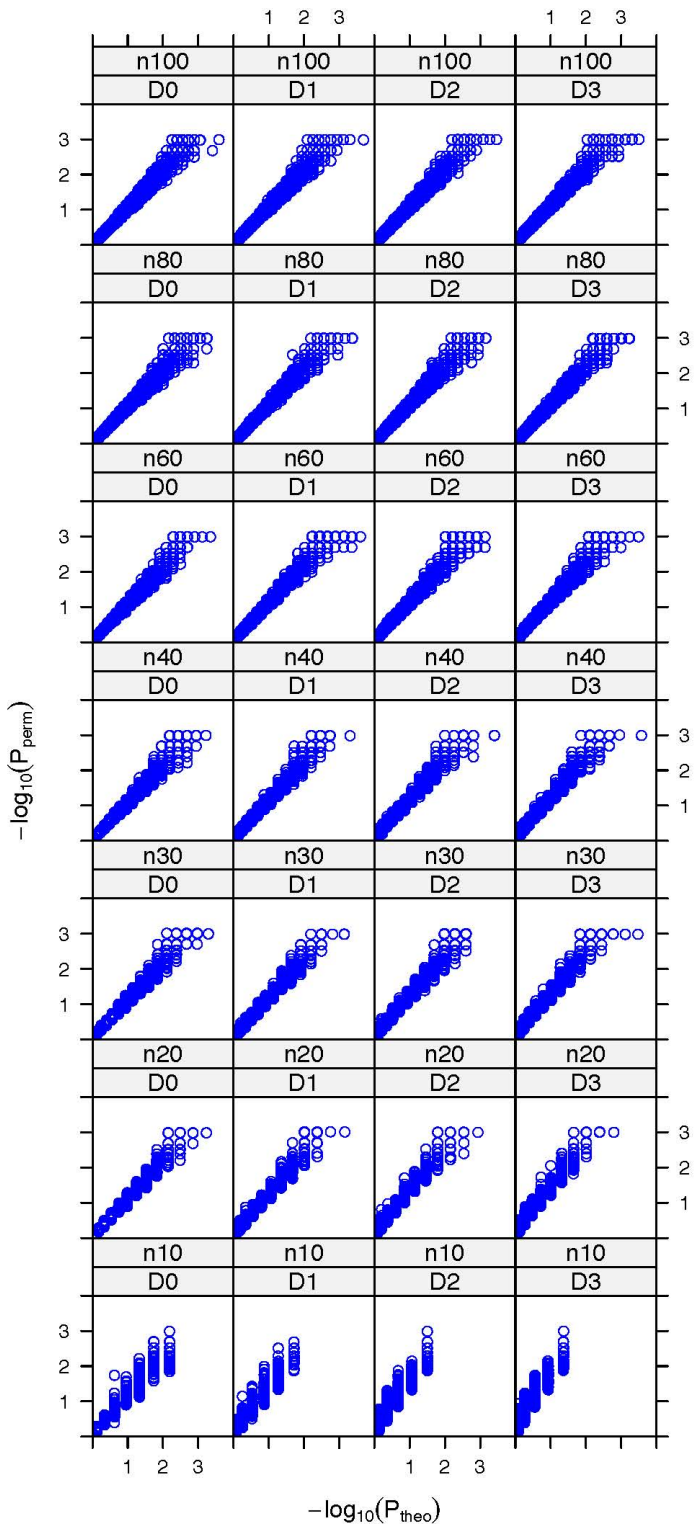

Figure B.2: Local shape analysis P_{theo} vs P_{perm} for 10,000 pairs simulated data. Columns D0 to D3 are for $D = 0, 1, 2, 3$. Rows n10 to n100 are for $n = 10, 20, 30, 40, 60, 80, 100$.

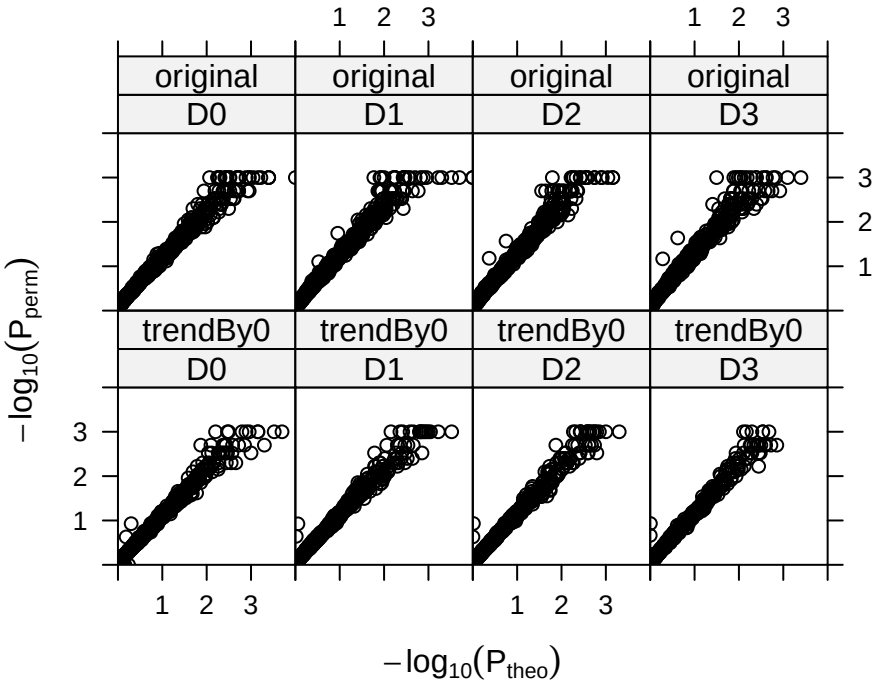

Figure B.3: P_{theo} and P_{perm} comparison for all-to-all pairwise local similarity ('original') and shape ('trendBy0') analysis of 40 abundant OTUs from SPOT dataset. Columns D0 to D3 are for $D = 0, 1, 2, 3$, respectively.

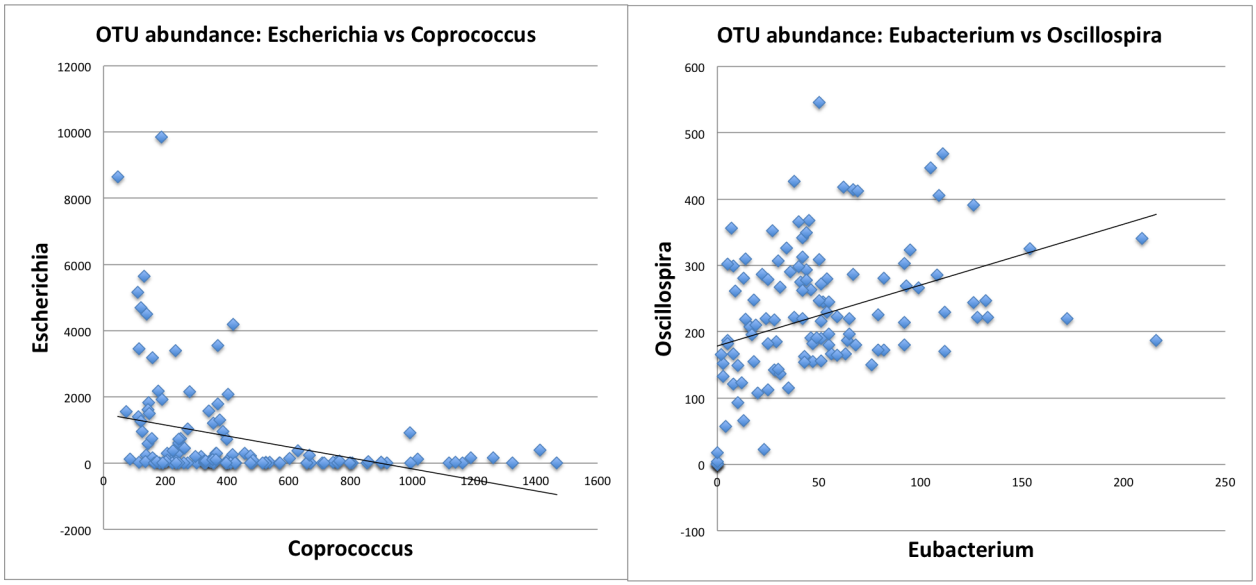

Figure B.4: Examples of significant local associations from 'F4' feces sample of MPH dataset. Profiles are shifted to synchronize the co-occurrence according to local similarity analysis. (left) *Coprococcus* and *Escherichia* (LS=-0.3179, P=0.0002; r=-0.3314, P=0.0001); (right) *Eubacterium* and *Oscillospira* (LS=0.3862, P=0.0001; r=0.3525, P=0.0001)

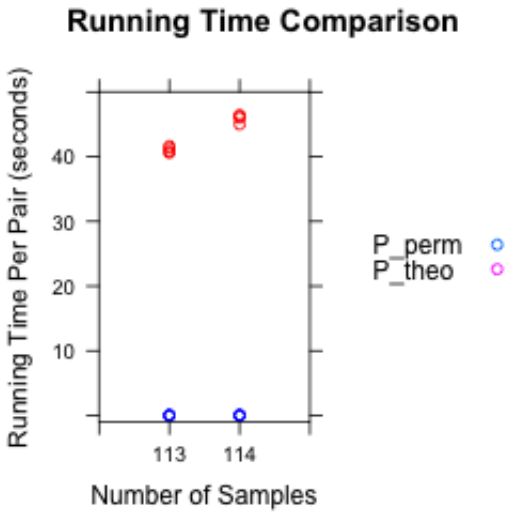

Figure B.5: Running time comparison for example real dataset computation. Note that the constant computation time using the theoretical approach that is independent of sample size as compared to sample-size and precision dependent computation time of permutation approaches.

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.1815	0.0491	0.1242	0.0921	0.0986	0.1092	0.1439	0.1308	0.1314	0.1371	0.1325	0.1511	0.1422
2.2	0.1111	0.0491	0.0677	0.0613	0.0680	0.0599	0.0832	0.0801	0.0859	0.0735	0.0770	0.0871	0.0888
2.4	0.0656	0.0132	0.0367	0.0353	0.0430	0.0420	0.0492	0.0487	0.0442	0.0486	0.0517	0.0509	0.0508
2.6	0.0373	0.0053	0.0168	0.0211	0.0158	0.0207	0.0266	0.0207	0.0288	0.0252	0.0283	0.0275	0.0234
2.8	0.0204	0.0053	0.0094	0.0057	0.0106	0.0098	0.0137	0.0125	0.0134	0.0112	0.0141	0.0140	0.0122
3.0	0.0108	0.0000	0.0039	0.0027	0.0034	0.0057	0.0069	0.0080	0.0075	0.0067	0.0066	0.0074	0.0065
3.2	0.0055	0.0000	0.0016	0.0013	0.0021	0.0026	0.0032	0.0044	0.0030	0.0032	0.0035	0.0041	0.0033
3.4	0.0027	0.0000	0.0002	0.0007	0.0009	0.0009	0.0008	0.0020	0.0016	0.0016	0.0013	0.0013	0.0016
3.6	0.0013	0.0000	0.0002	0.0000	0.0004	0.0005	0.0004	0.0009	0.0008	0.0005	0.0008	0.0007	0.0010
3.8	0.0006	0.0000	0.0000	0.0000	0.0001	0.0003	0.0001	0.0004	0.0001	0.0001	0.0005	0.0004	0.0002
4.0	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0001	0.0001	0.0002	0.0002	0.0002
4.2	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0000	0.0002	0.0001
4.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
4.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
4.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.5: Theoretical approximation for local shape analysis p-values versus the simulated probability $P(LS(D)/\sqrt{1.25n} \geq x)$. The theoretical approximate probability based on equation 2.3 with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{1.25n} \geq x$ is given in the 3rd to the 14th columns. $\mathbf{D = 0.}$

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.4516	0.0833	0.2633	0.2173	0.2282	0.2753	0.3391	0.2938	0.3116	0.3099	0.3124	0.3556	0.3352
2.2	0.2977	0.0833	0.1596	0.1415	0.1557	0.1512	0.2142	0.1929	0.2122	0.1799	0.1889	0.2263	0.2154
2.4	0.1841	0.0261	0.0837	0.0874	0.1073	0.1081	0.1255	0.1211	0.1111	0.1209	0.1295	0.1361	0.1327
2.6	0.1077	0.0038	0.0463	0.0496	0.0461	0.0549	0.0688	0.0570	0.0681	0.0633	0.0690	0.0795	0.0629
2.8	0.0601	0.0038	0.0209	0.0144	0.0293	0.0252	0.0368	0.0321	0.0317	0.0285	0.0353	0.0445	0.0322
3.0	0.0320	0.0000	0.0081	0.0073	0.0106	0.0169	0.0208	0.0181	0.0180	0.0183	0.0178	0.0231	0.0173
3.2	0.0164	0.0000	0.0031	0.0034	0.0059	0.0052	0.0105	0.0095	0.0084	0.0073	0.0087	0.0108	0.0079
3.4	0.0081	0.0000	0.0004	0.0015	0.0013	0.0018	0.0027	0.0022	0.0051	0.0046	0.0036	0.0039	0.0036
3.6	0.0038	0.0000	0.0004	0.0003	0.0007	0.0007	0.0021	0.0012	0.0024	0.0017	0.0021	0.0018	0.0017
3.8	0.0017	0.0000	0.0000	0.0000	0.0003	0.0000	0.0008	0.0004	0.0013	0.0002	0.0008	0.0005	0.0007
4.0	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0001	0.0006	0.0000	0.0004	0.0002	0.0001
4.2	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0001	0.0001	0.0000	0.0003	0.0001	0.0000
4.4	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000
4.6	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
4.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.6: Theoretical approximation for local shape analysis p-values versus the simulated probability $P(LS(D)/\sqrt{1.25n} \geq x)$. The theoretical approximate probability based on equation 2.3 with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{1.25n} \geq x$ is given in the 3rd to the 14th columns. **D = 1.**

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.6326	0.1180	0.3762	0.3057	0.3305	0.3924	0.4665	0.4317	0.4601	0.4528	0.4621	0.5079	0.4778
2.2	0.4452	0.1180	0.2275	0.1984	0.2337	0.2262	0.3020	0.2930	0.3275	0.2747	0.2984	0.3396	0.3221
2.4	0.2876	0.0308	0.1268	0.1236	0.1575	0.1684	0.1814	0.1897	0.1850	0.1866	0.2142	0.2080	0.2038
2.6	0.1730	0.0054	0.0631	0.0765	0.0671	0.0866	0.0993	0.0819	0.1153	0.0967	0.1201	0.1232	0.1050
2.8	0.0981	0.0054	0.0301	0.0239	0.0422	0.0398	0.0530	0.0471	0.0552	0.0461	0.0628	0.0667	0.0584
3.0	0.0528	0.0000	0.0120	0.0134	0.0138	0.0262	0.0283	0.0258	0.0304	0.0281	0.0297	0.0350	0.0314
3.2	0.0272	0.0000	0.0044	0.0070	0.0074	0.0108	0.0137	0.0135	0.0123	0.0124	0.0148	0.0166	0.0171
3.4	0.0134	0.0000	0.0002	0.0031	0.0025	0.0033	0.0043	0.0049	0.0061	0.0063	0.0070	0.0060	0.0090
3.6	0.0063	0.0000	0.0002	0.0005	0.0013	0.0011	0.0026	0.0027	0.0019	0.0024	0.0040	0.0030	0.0046
3.8	0.0029	0.0000	0.0000	0.0002	0.0005	0.0005	0.0008	0.0013	0.0011	0.0008	0.0016	0.0018	0.0016
4.0	0.0013	0.0000	0.0000	0.0000	0.0002	0.0001	0.0003	0.0006	0.0004	0.0005	0.0007	0.0007	0.0007
4.2	0.0005	0.0000	0.0000	0.0000	0.0002	0.0000	0.0001	0.0002	0.0001	0.0000	0.0003	0.0004	0.0002
4.4	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0002	0.0004	0.0002
4.6	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0001
4.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
5.0	0.0000	0.0001	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.7: Theoretical approximation for local shape analysis p-values versus the simulated probability $P(LS(D)/\sqrt{1.25n} \geq x)$. The theoretical approximate probability based on equation 2.3 with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{1.25n} \geq x$ is given in the 3rd to the 14th columns. **D = 2.**

		The number of time points n											
x	Theory	10	20	30	40	60	80	100	120	140	160	180	200
2	0.7539	0.1128	0.4443	0.3969	0.4194	0.4821	0.5824	0.5342	0.5639	0.5610	0.5598	0.6095	0.5956
2.2	0.5616	0.1128	0.2694	0.2690	0.3001	0.2906	0.3933	0.3684	0.4121	0.3551	0.3669	0.4284	0.4173
2.4	0.3779	0.0308	0.1480	0.1650	0.2068	0.2163	0.2490	0.2423	0.2356	0.2496	0.2633	0.2737	0.2732
2.6	0.2336	0.0044	0.0739	0.0960	0.0847	0.1126	0.1441	0.1178	0.1550	0.1362	0.1525	0.1643	0.1405
2.8	0.1346	0.0044	0.0351	0.0325	0.0528	0.0557	0.0744	0.0688	0.0708	0.0664	0.0837	0.0938	0.0806
3.0	0.0732	0.0000	0.0147	0.0161	0.0184	0.0359	0.0379	0.0379	0.0413	0.0408	0.0410	0.0493	0.0438
3.2	0.0379	0.0000	0.0060	0.0073	0.0105	0.0155	0.0183	0.0193	0.0166	0.0184	0.0190	0.0250	0.0237
3.4	0.0187	0.0000	0.0006	0.0038	0.0030	0.0064	0.0049	0.0072	0.0090	0.0102	0.0082	0.0083	0.0108
3.6	0.0089	0.0000	0.0006	0.0009	0.0013	0.0020	0.0038	0.0035	0.0039	0.0043	0.0043	0.0036	0.0055
3.8	0.0040	0.0000	0.0002	0.0003	0.0006	0.0009	0.0015	0.0016	0.0018	0.0022	0.0018	0.0020	0.0022
4.0	0.0018	0.0000	0.0001	0.0002	0.0002	0.0001	0.0004	0.0002	0.0006	0.0008	0.0009	0.0009	0.0006
4.2	0.0007	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0000	0.0004	0.0003	0.0002	0.0005	0.0003
4.4	0.0003	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0002	0.0000	0.0001	0.0003	0.0002
4.6	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0002	0.0001
4.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.8: Theoretical approximation for local shape analysis p-values versus the simulated probability $P(LS(D)/\sqrt{1.25n} \geq x)$. The theoretical approximate probability based on equation 2.3 with $\sigma = 1$ is given in the 2nd column and the simulated probability that $LS(D)/\sqrt{1.25n} \geq x$ is given in the 3rd to the 14th columns. **D = 3.**

(a) original, D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	233	0
P_{perm} ≤0.05	19	48

(b) original, D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	225	0
P_{perm} ≤0.05	37	38

(c) original, D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	228	0
P_{perm} ≤0.05	36	36

(d) original, D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	228	0
P_{perm} ≤0.05	36	36

(e) trendBy0, D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	241	0
P_{perm} ≤0.05	6	53

(f) trendBy0, D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	244	0
P_{perm} ≤0.05	9	47

(g) trendBy0, D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	243	0
P_{perm} ≤0.05	19	38

(h) trendBy0, D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	240	0
P_{perm} ≤0.05	22	38

Table B.9: The comparison of significant gene pairs using P_{theo} and P_{perm} given type-I error 0.05 for local similarity ('original') and shape ('trendBy0') analysis of 25 randomly selected factors from the CDC dataset: (a-d), local similarity scores; (e-h), local shape scores; (a,e) D=0, (b,f) D=1, (c,g) D=2, (d,h) D=3. The total number of comparisons is 300.

(a) real, D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	488	0
P_{perm} ≤0.05	31	261

(b) real, D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	492	0
P_{perm} ≤0.05	47	241

(c) real, D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	505	0
P_{perm} ≤0.05	57	218

(d) real, D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	516	0
P_{perm} ≤0.05	53	211

(e) trendBy0, D=0	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	578	0
P_{perm} ≤0.05	20	182

(f) trendBy0, D=1	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	606	0
P_{perm} ≤0.05	26	148

(g) trendBy0, D=2	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	622	0
P_{perm} ≤0.05	23	135

(h) trendBy0, D=3	P_{theo} >0.05	P_{theo} ≤0.05
P_{perm} >0.05	627	0
P_{perm} ≤0.05	31	122

Table B.10: The comparison of significant OTU pairs using P_{theo} and P_{perm} given type-I error 0.05 for local similarity ('original') and shape ('trendBy0') analysis of 40 selected OTUs from SPOT dataset: (a-d), local similarity scores; (e-h), local shape scores; (a,e) D=0, (b,f) D=1, (c,g) D=2, (d,h) D=3. The total number of OTU pairs is 780.