Title

Generalized correlation measure using count statistics for gene expression data with ordered samples

Journal

Bioinformatics, 34(4)

ISSN

1367-4803

Authors

Wang, YX Rachel

Liu, Ke

Theusch, Elizabeth

et al.

Publication Date

2018-02-15

DOI

10.1093/bioinformatics/btx641

Peer reviewed

Bioinformatics, 34(4), 2018, 617–624
doi: 10.1093/bioinformatics/btx641
Advance Access Publication Date: 12 October 2017
Original Paper

OXFORD

Gene expression

Generalized correlation measure using count statistics for gene expression data with ordered samples

Y. X. Rachel Wang[1], Ke Liu[2], Elizabeth Theusch[3], Jerome I. Rotter[4], Marisa W. Medina[3], Michael S. Waterman[5,*] and Haiyan Huang[2,*]

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia, [2]Department of Statistics, University of California, Berkeley, CA 94720, USA, [3]Children's Hospital Oakland Research Institute, Oakland, CA 94609, USA, [4]The Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, LABioMed at Harbor-UCLA Medical Center, Torrance, CA 90502, USA and [5]Molecular and Computational Biology, University of Southern California, CA 90089, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on March 13, 2017; revised on July 30, 2017; editorial decision on October 7, 2017; accepted on October 11, 2017

Abstract

Motivation: Capturing association patterns in gene expression levels under different conditions or time points is important for inferring gene regulatory interactions. In practice, temporal changes in gene expression may result in complex association patterns that require more sophisticated detection methods than simple correlation measures. For instance, the effect of regulation may lead to time-lagged associations and interactions local to a subset of samples. Furthermore, expression profiles of interest may not be aligned or directly comparable (e.g. gene expression profiles from two species).

Results: We propose a count statistic for measuring association between pairs of gene expression profiles consisting of ordered samples (e.g. time-course), where correlation may only exist locally in subsequences separated by a position shift. The statistic is simple and fast to compute, and we illustrate its use in two applications. In a cross-species comparison of developmental gene expression levels, we show our method not only measures association of gene expressions between the two species, but also provides alignment between different developmental stages. In the second application, we applied our statistic to expression profiles from two distinct phenotypic conditions, where the samples in each profile are ordered by the associated phenotypic values. The detected associations can be useful in building correspondence between gene association networks under different phenotypes. On the theoretical side, we provide asymptotic distributions of the statistic for different regions of the parameter space and test its power on simulated data.

Availability and implementation: The code used to perform the analysis is available as part of the Supplementary Material.

Contact: msw@usc.edu or hhuang@stat.berkeley.edu

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

1 Introduction

Understanding complex regulatory relationships between genes is one of the central themes of systems biology. As high-throughput technologies continue to generate large-scale gene expression datasets, developing efficient computational and statistical tools to infer or reconstruct gene interactions remains a highly relevant area of

research. It is commonly assumed that co-regulation relationships can be partially deduced from expression correlation patterns. For example, when gene expression levels are measured in a time-course experiment, similar expression profiles between gene pairs suggest possible activation relationships, while inverted profiles may imply inhibition. Extracting meaningful patterns from these expression profiles is often the first step toward analyzing functional groupings of genes, annotating unknown genes and reconstructing gene regulatory networks.

Treating the problem as that of detecting statistical correlation, Pearson's correlation (PC) has been one of the most widely used measures for finding gene pairs with similar expression profiles (Eisen *et al.*, 1998; Stuart *et al.*, 2003; Wolfe *et al.*, 2005). Beyond assessing linear dependence, another class of methods based on mutual information measuring general statistical dependence has also been extensively used (Basso *et al.*, 2005; Daub *et al.*, 2004; Margolin *et al.*, 2006; Steuer *et al.*, 2002). For time-course data, techniques in times series analysis (e.g. time-frequency analysis) have been applied to improve the sensitivity of similarity measures (Feng *et al.*, 2004; Ramoni *et al.*, 2002), often assuming explicit models for generating the observed data.

Although an extensive collection of methods exists for measuring global association between gene pairs, co-regulation and consequently gene expression dependence may only exist across a subset of the experimental conditions. Pei *et al.* (2014) addressed this problem by introducing a spline regression model and a penalized PC score. Non-parametric methods comparing local expression patterns were introduced in Roy *et al.* (2014) and Wang *et al.* (2014), with the latter method applicable to both time series and more general datasets. Biclustering offers an alternative line of approach by simultaneously finding subsets of genes and subsets of experimental conditions under which association patterns exist. However, biclustering methods often come at a high computational cost and are formulated using specific generative models such as the additive model and the multiplicative model (Cheng and Church, 2000; Gao *et al.*, 2012; Hochreiter *et al.*, 2010; Lazzeroni and Owen, 2002). Although one advantage of biclustering is it outputs groups of genes with potentially related functions, pairwise similarity measures provide a more direct way to rank and compare gene pairs in terms of their functional relatedness.

Another prevalent feature of time series data is the presence of time lags between association patterns, reflecting the fact that regulation may take effect after a time delay. Early work to incorporate this feature includes the time-shifted PC proposed in Kato *et al.* (2001). Motivated by the large body of work in sequence alignment, Kwon *et al.* (2003) summarized expression patterns as character strings and used the Needleman-Wunsch algorithm to compute optimal global alignment scores for gene pairs. Instead of matching pairs of time points, another class of methods known as dynamic time warping (DTW) aligns two time series globally based on Euclidean distance minimization. Originally developed for speech recognition, DTW has been widely applied in comparative analysis of temporal gene expression data from different species (Aach and Church, 2001; Goltsev and Papatsenko, 2009; Smith *et al.*, 2008; Yuan *et al.*, 2011). However, the DTW approach can only be applied to genes with similar temporal profiles in order to correctly estimate time shifts, thus limiting its use mostly to gene pairs filtered by a correlation test or orthologous genes in cross-species comparison.

The above approaches for handling time shifts are designed to measure global associations. To detect local correlation patterns, one natural extension is to consider optimal local alignment of expression subsequences where optimality is formulated as maximal correlation or matching expression patterns (Balasubramaniyan *et al.*, 2005; Ji and Tan, 2005; Qian *et al.*, 2001). However, such methods may assign a low score when a gene pair has several segments of subsequences with strong correlation but the segments are not sufficiently long.

In this article, we develop a count statistic for measuring association between pairs of gene expression profiles with ordered samples (including time-course data), where (i) correlation may only exist locally in subsequences; (ii) the correlated subsequences can be separated by a position shift; (iii) the size of shifts may also differ from location to location—in this case the pair of expression profiles cannot be aligned globally on the same scale. The statistic sums up the number of subsequence pairs with matching ranks, thus has better sensitivity for detecting multiple but relatively short correlated segments. The statistic is simple and fast to compute. As the statistics is rank-based, it captures more information than methods that only consider the rise and fall of expression levels (Ji and Tan, 2005; Kwon *et al.*, 2003) while being nonparametric and flexible. We provide asymptotic analysis for different regions of the parameter space and examine the accuracy of the approximation by simulation. To test the performance of the statistic as a measure of local associations, we compare the power of the statistic with Local Similarity Analysis (LSA, Ruan *et al.*, 2006; Xia *et al.*, 2013) using simulated data. LSA is a similarity metric originally designed for studying interactions among microbial communities and detecting their temporal associations. Using PC as a local similarity measure, LSA belongs to the general class of methods that tries to align the most similar subsequences. We show that our statistic has a better power when there are multiple but short correlated subsequences.

In general, our statistic is useful for identifying associations between expression profiles consisting of ordered samples, when the samples are not aligned and contain local association patterns possibly separated by gaps. In this paper, we demonstrate the utility of our statistic in two different applications. In the first application, we applied the statistic to compare developmental gene expression levels in fly and worm. We demonstrate the statistic can be used as a measure of association between a fly gene and a worm gene. In addition, the computation of our statistic naturally leads to direct alignment of subsequences indicating where strong correlations exist and allows one time point in one sequence to be mapped to possibly multiple time points in the other. This feature makes the statistic a suitable tool for visualizing the temporal alignment and studying correspondence between different time points in a cross-species comparison. Previously the alignment between developmental stages was done by identifying enrichment of orthologous genes Li *et al.* (2014); last we show this can be achieved in a more general way and extend the analysis to include general, non-orthologous gene pairs.

As a second application, we consider calculating associations between gene expression profiles coming from two phenotypic conditions. Taking RNA-seq data from the Cholesterol and Pharmacogenetics (CAP) clinical trial (Simon *et al.*, 2006), we ordered the samples by the individuals' low-density lipoprotein cholesterol (LDLC) levels and identified a high and a low LDLC group. We next used our statistic to compute gene cross-correlations between these two phenotypic groups for a set of cholesterol metabolism genes. In this case, the samples in the two groups came from different individuals and hence cannot be aligned directly. Given the high degrees of coexpression between the gene pairs within the high and low LDLC groups themselves, we expect a sensitive statistic would find more cross-correlations between the two LDLC groups. We show our method is more sensitive than global correlation measures and the local correlation measure LSA. This technique makes it

possible to build correspondence between gene association networks constructed under high and low LDLC levels.

2 Materials and methods

2.1 Defining the correlation measures

For two gene expression profiles with ordered samples (e.g. time-course data) $\mathbf{x} = (x_1, \ldots, x_m)$, $\mathbf{y} = (y_1, \ldots, y_n)$, we define a correlation measure for capturing local association patterns in $\mathbf{x}$ and $\mathbf{y}$. To motivate the definition, consider comparing gene expression levels measured at different developmental stages in the two species *Drosophila melanogaster* and *Caenorhabditis elegans*. Figure 1 shows the estimated mapping between the embryonic stages (Li *et al.*, 2014). Although the two timelines cannot be aligned globally with a simple time shift, subsequences of time points located at different positions can be matched. This suggests functionally related genes in these two species can have correlated expression patterns between matched subsequences, thus their overall correlation can be calculated. However, in practice the stage correspondence is usually unknown a priori. In addition, the matching local regions are separated by time shifts of different lengths, making it hard for existing methods to detect these local patterns. Motivated by the above observation, we design a statistic to compare subsequences of expression levels starting at different positions in the time series (up to some maximal time shift).

We measure the association between a pair of subsequences by comparing their rank patterns. Define a rank function $\phi(\cdot)$, which describes the rank profile of a vector and returns the indices of elements in the vector after they have been sorted in an increasing order. For example, for the vector (1, 5, 3), ϕ returns (1, 3, 2). For two subsequences of length k starting at position i and j in $\mathbf{x}$ and $\mathbf{y}$ respectively, we check whether their rank patterns are identical or reversed using the indicators

$$\mathbb{I}^{+}_{(i,j)} = \mathbb{I}\big(\phi(x_i, \ldots, x_{i+k-1}) = \phi(y_j, \ldots, y_{j+k-1})\big),$$

$$\mathbb{I}^{-}_{(i,j)} = \mathbb{I}\big(\phi(x_i, \ldots, x_{i+k-1}) = \phi(-y_j, \ldots, -y_{j+k-1})\big).$$

The sum of the indicators over all possible position pairs i and j measures the extent of co-variation between $\mathbf{x}$ and $\mathbf{y}$. For convenience denote $\bar{m} = m - k + 1$ and $\bar{n} = n - k + 1$. The correlation measures are defined as

$$V^{+} = \sum\nolimits_{v \in \mathcal{S}} \mathbb{I}^{+}_{v}, \quad V^{-} = \sum\nolimits_{v \in \mathcal{S}} \mathbb{I}^{-}_{v}, \quad V = V^{+} + V^{-},$$

where $v = (i, j)$ and $\mathcal{S} = \{v = (i, j) \mid 1 \le i \le \bar{m}, 1 \le j \le \bar{n}, |i - j| \le d\}$ for some maximum shift d. For example, given $\boldsymbol{x} = (1.2, 2.3, 3.5, 4.1, 5.8)$ and $\boldsymbol{y} = (3.1, 2.4, 0.9, 4.3, 5.5)$, then for $k = 3$ and maximal time shift $d = 1$, $\mathbb{I}^{+}_{(2,3)} = 1$, $\mathbb{I}^{+}_{(3,3)} = 1$, making $V^{+} = 2$; $\mathbb{I}^{-}_{(1,1)} = 1$, $\mathbb{I}^{-}_{(2,1)} = 1$, making $V^{-} = 2$. Overalls $V = 4$. We note that V generalizes the measure W_1 in Wang *et al.* (2014), which only compares subsequences starting at the same positions in a pair of expression profiles. V accounts for possible time shifts in gene interactions, and d is the maximum time shift allowed for interactions to take place. With this generalization, V can be used in situations where the expression profiles of interest are not directly aligned. V^{+} and V^{-} are one-sided counts to be used when the local associations are believed to be all positive or all negative.

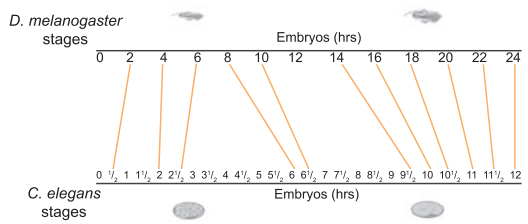

Fig. 1. Figure adapted with permission from Figure 5B in Li *et al.* (2014). Correspondence between embryonic stages in *D. melanogaster* and *C. elegans*

2.2 Asymptotic distributions

The measures defined above can be used directly to rank gene pairs in terms of their association strength. With appropriate assumptions on the distribution of $\mathbf{x}$ and $\mathbf{y}$, we can further characterize the asymptotic distribution of V on a pair of independent sequences, which allows us to approximate P-values for long series. Depending on how k (subsequence length) and d (maximal shift) varies in the limit, we have a normal and a Poisson limiting regime for the distribution of V. For notational convenience we consider the case $m = n$, although the conclusions remain true as long as m grows at the same rate as n. The exact statement of the assumptions and proofs can be found in the Supplementary Material. Although the techniques used are similar to Wang *et al.* (2014), the inclusion of the time lag d complicates the theoretical analysis.

2.2.1 Normal limit

Define the normalized statistic $\bar{V} = (V - \mu_n)/\sigma_n$. Here $\mu_n = \mathbb{E}(V) = \frac{2}{k!} n_{k,d}$ with $n_{k,d} = (2d + 1)\bar{n} - d(d + 1)$ (see Supplementary Material), which can be computed explicitly. The variance $\sigma_n^2 = \mathrm{Var}(V)$ needs to be estimated by Monte Carlo simulations. In particular, one can simulate independent pairs of iid sequences (or exchangeable sequences, see Assumption 2 in the Supplementary Material) and approximate the variance of the statistic using sample variance. As $n \to \infty$, the limiting distribution of $\bar{V}$ is standard normal, i.e.

$$\bar{V} \xrightarrow{D} N(0, 1) \tag{1}$$

when $\mu_n \to \infty$ and either of the following holds:

- k fixed, $d^3/n \to 0$;
- d fixed, $k/(\log n)^{\alpha} \to 0$ for $\alpha < 1$.

In other words, the approximation holds for small values of d and k. As we will later demonstrate using simulated and real data, setting k to be approximately $\log n$ often produces ideal results. The choice of d depends more on specific applications and user knowledge of the size of the maximal time shift.

2.2.2 Poisson limit

As k increases, subsequences with exact matches become increasingly rare resulting in a Poisson limit regime. In this case, as $n \to \infty$, the distribution of V can be approximated by a Poisson random variable with equal mean. i.e.

$$d_{\mathrm{TV}}(V, Z) \to 0, \tag{2}$$

provided $\mu_n = O(1)$ and $d/k \to 0$, where $Z \sim \mathrm{Poisson}(\mu_n)$ and d_{TV} is the total variation distance. Roughly the Poisson approximation holds for large k and small d. In real applications, the Poisson approximation is less useful than the normal one, since finding exact matches of subsequences becomes increasingly rare and less practical with large k.

2.3 Application to simulated data

Using simulated data, we compare our method with the LSA (Ruan *et al.*, 2006; Xia *et al.*, 2013), which is one of the recent methods for measuring local correlation with source code available. LSA uses dynamic programing to align subsequences with a large PC in two time series. We generated pairs of times series with 500 time points containing locally correlated regions masked by time delays and other noisy regions. To generate **x**, we obtained x_i from a AR(1) time series with coefficient 0.1. Next **y** was constructed as follows. For a signal segment s, a set of indices $\mathcal{I}_s$ associated with **x**, and a time delay d_s, we have

$$y_{i+d_s} = x_i + e_i \qquad \text{for } i \in \mathcal{I}_s$$

with e_i being independent normal noise with standard deviation σ_e. Outside the signal regions, y_i followed an independent AR(1) time series with coefficient -0.2. For example, scenario (ii) in Table 2 contains one correlated signal segment with $s = 1$, $\mathcal{I}_1 = \{201, \ldots, 250\}$, $|\mathcal{I}_1| = 50$, $d_1 = 5$, $\sigma_e = 0.3$.

For each simulated dataset, we computed the normalized statistic $\bar{V}$ and approximated the P-values using standard normal distribution. The method provided in Xia *et al.* (2013) was used to approximate P-values for the LSA scores. We compared the power of the two methods rejecting the null hypothesis of independence at 5% and 1% significance levels for different values of k and d on datasets with different numbers of signal segments, segment lengths, and noise levels. Under all scenarios, $d_s \leq 5$ for all s.

2.4 Application to cross-species comparison of gene expression in developmental stages

We illustrate the use of V in a cross-species comparison of gene expression levels for two model organisms *D. melanogaster* (fly) and *C. elegans* (worm). These two species are evolutionarily distant and have morphologically distinct developmental stages. Since the alignment of their developmental timelines is unknown, genome-wide comparison of gene expression across multiple stages is a challenging problem. Li *et al.* (2014) provided the first comprehensive study using modENCODE (Celniker *et al.*, 2009) data and orthologous gene pairs to assess the similarity between different stages. We show similar results can be achieved without limiting the analysis to orthologous gene pairs.

We used the time-course RNA-seq data in Li *et al.* (2014). The fly time-course data consists of 30 time points (embryos, L1–L3 larvae, pupae, male and female adults); the worm time-course data contains 35 time points (embryos, L1–L4 larval, young adults, adults and dauer). Since male and female adults represent two parallel stages of development in fly (Li *et al.*, 2014, Fig. 1A), we further split the fly data into dm_male (27 time points) and dm_female (27 time points). Similarly, as the dauer stage in worm is an alternative development prior to the L4 larval stage (Li *et al.*, 2014, Fig. 1B), we split the worm data into ce_nd (32 time points, no dauer stage) and ce_da (33 time points, with dauer stage).

Systematic comparisons of gene expression changes across different time points between different species provide valuable information on fundamental patterns and dynamic features conserved or altered through evolution. Although it is natural to assume time stages sharing similar transcriptional activities are related, the challenge remains in identifying patterns of similarity between gene expression profiles when the time points are not aligned. Li *et al.* (2014) circumvented this difficulty by assuming orthologous genes have similar transcriptional characteristics; we showed such characteristics can be directly measured using our statistic. To this end, we first demonstrated our statistic is informative for measuring association between orthologous genes, which tend to be more correlated than non-orthologous ones. Then by identifying the locations of correlated subsequences in orthologous gene pairs, we constructed stage correspondence maps similar to the ones in Li *et al.* (2014). Finally, we were able to show this analysis can be extended to general gene pairs, thus removing the dependence on extra annotation information and making the approach applicable to less well-studied species.

2.4.1 Filtering gene pairs

Using the list of orthologous genes in Li *et al.* (2014), we first filtered them and selected those that (i) have FPKM (Fragments Per Kilobase of transcript per Million mapped reads) $\geq$1.0 across all time points and (ii) have expression levels not always ranked in the top 30% or the bottom 30% of all the genes satisfying (i), leaving us a total of 3116 fly genes and 3133 worm genes. 2761 pairs of these are orthologous. We note that criterion (ii) was used to select genes with a reasonable amount of variation across all the time points.

In order to extend the same analysis to general gene pairs, we used the same filtering criterion as earlier, but this time applied to all the available genes, to obtain 5379 (fly) $\times$ 5368 (worm) general gene pairs. Similar to the above, filtering was applied only to remove genes with small or large variance; orthology information was not used in this case.

2.4.2 Correlation calculation between orthologous gene pairs

Since we expect orthologous genes to be functionally related, we next tested whether our statistic can be used to show orthologous gene pairs tend to be more correlated than non-orthologous ones. We used V^+ to measure how positively correlated the orthologous pairs are compared with the rest of the gene pairs. Using $k = 5$ and maximum d, we computed V^+ for every pair of fly and worm genes in the 3116 $\times$ 3133 matrix. For comparison, we also calculated all the pairwise scores using LSA, again making the measure one-sided by only calculating the positive scoring matrix. Using the scores to rank all the gene pairs in a descending order, we compared the two methods by counting the number of orthologous pairs included in the top ranked pairs.

2.4.3 Construction of stage correspondence maps

Here we introduce an extended application of our statistic V^+ to identify correspondence between different time points. We considered building an $\bar{m} \times \bar{n}$ correspondence matrix, where each entry represents a possible match between time points in the two species. The matches can be obtained from the local similarity profiles contained in the computation of V^+. We first explored this idea using orthologous pairs and then extended the analysis to general gene pairs.

We set $k = 5$ and d to its maximum value to allow for all possible time shifts. In the $\bar{m} \times \bar{n}$ correspondence matrix, for every orthologous pair, the entries $\{(i,j), \ldots, (i+4, j+4)\}$ received count 1 if $\mathbb{I}^+_{(i,j)} = 1$, indicating the existence of local similarity between all five pairs of stages. The total counts were thresholded at the 85% quantile and the final binary heatmaps allowed us to visualize matching between different stages in the two species.

Next we extended our analysis of local similarity to the 5379 (fly) $\times$ 5368 (worm) general gene pairs. We used the V^+ scores to select gene pairs that are likely to be related in the two species. In the 5379 $\times$ 5368 V^+ score matrix, we extracted gene pairs with the highest scores sequentially, each time removing from the matrix the row (the fly gene) and the column (the worm gene) that had been chosen to avoid repetition. For every dataset combination (e.g.

dm_male with ce_nd), about 1600 gene pairs were chosen. The same count matrix for stage correspondence, as what we constructed for the orthologous genes pairs, was computed as described before. The thresholding value was set at the 85% quantile of all the counts.

2.5 Identifying interactions between lipid genes relative to phenotypic variation

To further demonstrate the general utitlity of our method, we considered using it to find correspondence between two gene association networks built under different conditions (e.g. two different phenotypes). Within each network, the edge weights are assigned using pairwise gene associations. Similarly, correspondence between the two networks can be found by measuring association between two genes, one from each network. Identifying such cross-network associations (and thus functional correspondence or relevance) is interesting as it would help better understand the functional relevance and difference of gene networks across different biological conditions. Traditional measures are usually not applicable to assess such associations. Our method is applicable, because (i) it can measure similarity between two genes even when their expression profiles consist of samples not directly aligned; (ii) it considers local similarity patterns, which are more likely to exist than global trends when the samples are not directly comparable.

We applied the statistic to an RNA-seq dataset consisting of measurements from ∼400 lymphoblastoid cell lines from participants of the CAP clinical trial (Simon *et al.*, 2006). Plasma lipid data and clinical covariates (age, sex etc.) are also available in the cell line donors. We focused on 21 highly coexpressed genes (specific to lymphoblastoid cell lines, K.Liu, personal communication) that primarily represent genes in the cholesterol synthesis pathway, and examined their pairwise relationships relative to the plasma LDLC levels in the individuals. We ordered the individuals by their LDLC levels. Treating the LDLC levels as time points, we compared the cross-correlations between 100 individuals with the lowest LDLC and 100 individuals with the highest LDLC for the 21 genes. Since in this case the 'time points' (or individuals) are not aligned, the cross-correlations are difficult to detect using traditional measures.

2.5.1 Data normalization and computation of correlation measures

The whole RNA-seq dataset was normalized using DESeq2 (Love *et al.*, 2014) and adjusted for potential confounders using probabilistic estimation of expression residual (Stegle *et al.*, 2012). The LDLC levels were adjusted for clinical covariates using regression. More detailed descriptions of the expression quantification, normalization, and adjustment of LDLC can be found in the Supplementary Material. We calculated PC, the LSA score and the $\bar{V}$ score for the 21 cholesterol synthesis genes both between the two LDLC sets and within the sets themselves, and computed their *P*-values.

3 Results

3.1 Agreement with asymptotic distributions

We first checked the accuracy of asymptotic approximation for different regions of d and k. Figure 2 shows the empirical quantiles of simulated $\bar{V}$ for different parameter settings. For each set, 1000 random permutations were simulated from which $\bar{V}$ was calculated. In Figure 2a, the distribution starts to deviate as k increases for fixed d. Similarly in Figure 2b, for fixed k larger d leads to more deviation from the normal, although the difference is less obvious. This agrees with our statement that the normal approximation (1) works well for small values of k and d.

To test the agreement with the Poisson limit, Table 1 shows the *P*-values obtained from the Kolmogorov-Smirnov test for different parameter settings. The test measures the distance between the empirical distributions of V and simulated Poisson(μ_n) random variables, recalling that μ_n is the expectation which can be explicitly calculated. For larger values of k, we expect the empirical distribution to move into the Poisson regime and the *P*-values to become larger. For the n and k values shown, smaller d leads to better agreement with the Poisson distribution since we require the ratio d/k to be small in (2).

3.2 Comparison with LSA on simulated data

Table 2 compares the power of $\bar{V}$ and LSA rejecting at 5% and 1% significance levels. 100 datasets were generated for each scenario following the description in the last section. When the two time series are independent, both methods have rejection rates around 0.05 as expected. Comparing scenarios (ii) and (iii), $\bar{V}$ has better power than LSA when the noise level is low. However, its power decreases with increasing noise level while LSA remains very robust. This is unsurprising noting that LSA uses PC as a measure of local signal strength, which is less stringent than $\bar{V}$. $\bar{V}$ requires local subsequences to have exact (or opposite) rank matches, which is equivalent to locally having a perfect Spearman's correlation. On the other hand, when the signal segment is separated into two parts in scenario (iv), $\bar{V}$ performs much better than LSA. Since both methods have conceptual connections to finding optimal local alignments in sequence matching, this scenario demonstrates that $\bar{V}$ is more robust to the insertion of noisy regions because unlike LSA which considers only the most optimally aligned subsequences, $\bar{V}$ includes contributions from all correlated subsequences. The same idea is reiterated in scenario (v), where there are three short signal segments separated by noise regions. Regarding the choice of the tuning parameters for $\bar{V}$, $k = 6$ (∼log 500) yields the best performance. Since the maximal lag between correlated segments is 5 in scenarios (ii–v), $d = 5$ leads to better power for both $\bar{V}$ and LSA.

3.3 Real data results

3.3.1 Cross-species comparison

3.3.1.1 Capturing correlation between orthologous genes. As expected, the orthologous pairs tend to have higher V^+ scores, suggesting that local correlations can be used as complementary evidence to define orthology. On the datasets dm_male and ce_nd, the V^+ values on 2761 pairs of orthologous gene pairs are statistically larger than the other non-orthologous pairs in the 3116×3133 matrix. The shift in distribution has a *P*-value which is ∼0 using the Wilcoxon rank sum test.

Ranking all the genes pairs in a descending order using V^+ and LSA, we compared the two methods by counting the number of orthologous pairs included in the top ranking pairs. As shown in Figure 3, V^+ consistently chooses more orthologous pairs than LSA. We note that although the proportion of identified orthologous gene pairs appears low (around 300/800 000) for both methods, given there exist only 2761 true positives among all 3116×3133 possible pairs, this proportion identified is highly statistically significant with a *P*-value of 1.7×10^{-6} using the binomial test. This suggests V^+ is informative for measuring orthology to some extent, although could be made more powerful if combined with other sources of information such as sequence similarity. The results from the other dataset combinations (e.g. dm_female with ce_da) are very similar and can be found in Section 2 of the Supplementary Material.

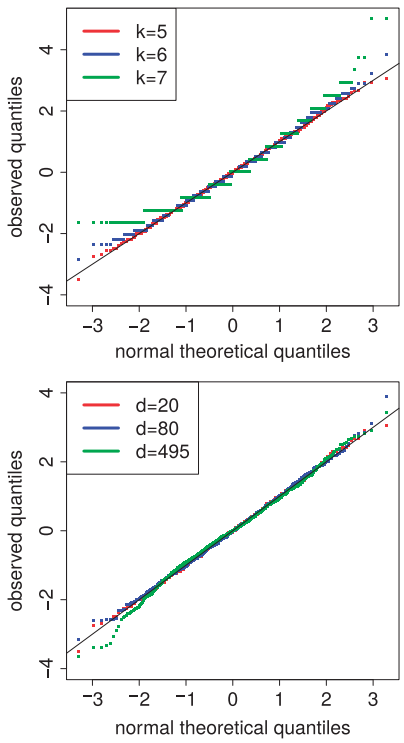

Fig. 2. Empirical quantiles for (a) $n = 500$, $d = 20$, varying k and (b) $n = 500$, $k = 5$, varying d from 1000 simulated random permutations

Table 1. *P*-values from the Kolmogorov-Smirnov test measuring the distance between the empirical distributions of V and simulated Poisson(μ_n) random variables (2000 simulations)

d	$n = 50, k = 7$	$n = 500, k = 8$	$n = 2000, k = 9$
1	0.056	0.118	0.452
2	0.002	0.097	0.427
3	0	0.001	0.187
4	0	0	0
5	0	0	0.06

3.3.1.2 Correspondence between developmental stages in fly and worm. Figure 4 shows the binary heatmaps constructed using the method described in the last section for each dataset using orthologous pairs. Similar to the results in Li *et al.* (2014), two parallel patterns exist between fly and worm from embryos to larvae, and between fly pupae and worm embryos. As explained in Li *et al.* (2014), the upper parallel line is consistent with the second wave of cell proliferation and differentiation in the fly life cycle. The dataset combination dm_female and ce_da is shown in Figure 4c. Additional correspondence exists between fly pupae and worm dauer. However, the mappings between fly female and the worm stages are missing due to edge effects. We note that since LSA looks for the optimal local alignment of subsequences, it is less straightforward to extract correspondence between all pairs of stages.

Extending the analysis to the general gene pairs chosen based on their V^+ scores, the same count matrices were calculated and thresholded. Figure 4b and d shows the stage mappings for the same datasets. It can be seen that the maps maintain the major features of their orthologous counterparts with some noise, indicating the main conclusions previously drawn on stage correspondence can be generalized without restricting the analysis to orthologous genes.

3.3.1.3 Cross-correlations between low and high LDLC samples. Our goal is to show $\bar{V}$ can be used to identify correspondence between two genes networks. In this dataset, all pairs in the 21 cholesterol metabolism gene set show very strong PC in both conditions. That is, the two networks are almost fully connected. Hence we expect in general that gene pairs across the two networks (each from one network) should be well connected, and we can use the number of correlated gene pairs found to evaluate the performance of different methods.

Table 3 compares the number of significant cross-correlations found between the low and high LDLC samples for the 21 genes. To compute the $\bar{V}$ scores, we set $k = 4$ [$\sim$log 100]; different lag values d were tested for both $\bar{V}$ and LSA. At both *P*-value cutoffs (0.05 and 0.01), $\bar{V}$ tends to detect more correlations than LSA. The correlations found by $\bar{V}$ at $k = 4, d = 10$ are plotted in a bipartite network in Supplementary Figure S3. We note that the most significant edges in this network (e.g. the top 20 edges with the smallest *P*-values)

Table 2. Fractions of *P*-values smaller than 0.05 and 0.01 out of 100 simulations when (i) x_i and y_i are independent; (ii) one signal segment of length 50, $\sigma_e = 0.3$; (iii) one signal segment of length 50, $\sigma_e = 0.5$; (iv) two signal segments each of length 25, $\sigma_e = 0.3$; (v) three signal segments each of length 25, $\sigma_e = 0.3$

		$\bar{V}$				LSA		
		$k = 5, d = 5$	$k = 6, d = 5$	$k = 6, d = 6$	$k = 6, d = 7$	$d = 5$	$d = 6$	$d = 7$
(i) Independent	($P < 0.05$)	0.02	0.04	0.05	0.04	0.04	0.03	0.03
	($P < 0.01$)	0.02	0.02	0.01	0.02	0.01	0.01	0.01
(ii) $s = 1$, $\lvert\mathcal{I}_s\rvert = 50$, $\sigma_e = 0.3$	($P < 0.05$)	0.41	0.67	0.57	0.59	0.43	0.43	0.41
	($P < 0.01$)	0.30	0.49	0.42	0.38	0.14	0.14	0.13
(iii) $s = 1$, $\lvert\mathcal{I}_s\rvert = 50$, $\sigma_e = 0.5$	($P < 0.05$)	0.21	0.29	0.24	0.24	0.41	0.39	0.36
	($P < 0.01$)	0.10	0.15	0.13	0.12	0.14	0.15	0.14
(iv) $s = 1, 2$, $\lvert\mathcal{I}_s\rvert = 25$, $\sigma_e = 0.3$	($P < 0.05$)	0.43	0.58	0.53	0.48	0.11	0.10	0.10
	($P < 0.01$)	0.23	0.44	0.38	0.32	0.01	0.01	0.02
(v) $s = 1, 2, 3$, $\lvert\mathcal{I}_s\rvert = 25$, $\sigma_e = 0.3$	($P < 0.05$)	0.58	0.85	0.83	0.79	0.53	0.42	0.39
	($P < 0.01$)	0.33	0.71	0.66	0.59	0.19	0.19	0.15

$d_s \leq 5$ for all s under all the scenarios.

could not be identified by PC using 0.05 as the *P*-value cutoff, highlighting the necessity to consider local correlation patterns in this case. Computing PC for these genes within the low and high LDLC samples themselves, we note that the correlations are uniformly high for all gene pairs, suggesting these genes, which all have functions related to cholesterol metabolism, have high levels of interaction and the cross-correlations found are reasonable. Finally, we observe that compared with PC, $\bar{V}$ finds much fewer significant interactions in the within-sample calculations. A similar observation holds for LSA. This suggests PC remains an ideal measure for detecting global association patterns when the samples are matched, but alternative measures targeting local correlations can provide complementary information when there is no clear alignment between the samples.

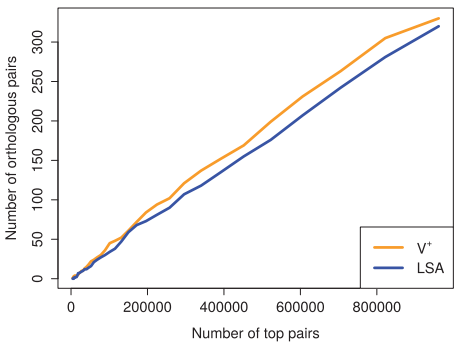

Fig. 3. Number of orthologous pairs included in top gene pairs ranked by V^+ (orange) and LSA (blue) (Color version of this figure is available at *Bioinformatics* online.)

4 Discussion

As the amount of high-throughput gene expression data continues to accumulate, developing appropriate statistical tools to extract meaningful association patterns remains an important problem and has broader applications in general data mining. For time-course expression data, the use of traditional correlation methods is often limited by the presence of time shifts between correlated time points and the fact that these association patterns may only exist locally. Many existing methods designed to address these problems have conceptual similarity to sequence alignment algorithms. Optimal local alignments of expression subsequences are found by maximizing some measure of association strength. However, such methods may not be well suited to detect patterns with multiple correlated segments, especially when these segments are relatively short in length. In this article, we propose a count statistic that sums up contribution from all subsequence pairs of length k having matching ranks separated by a maximal time shift d. We provide the asymptotic distributions of the statistic

Table 3. Number of significant correlations between low and high LDLC samples for different choices the lag parameter d and *P*-value cutoffs

		$d = 8$	$d = 10$	$d = 12$	$d = 14$
$P < 0.05$	$\bar{V}$	22	25	20	25
	LSA	14	22	19	18
$P < 0.01$	$\bar{V}$	5	5	3	5
	LSA	2	4	3	6

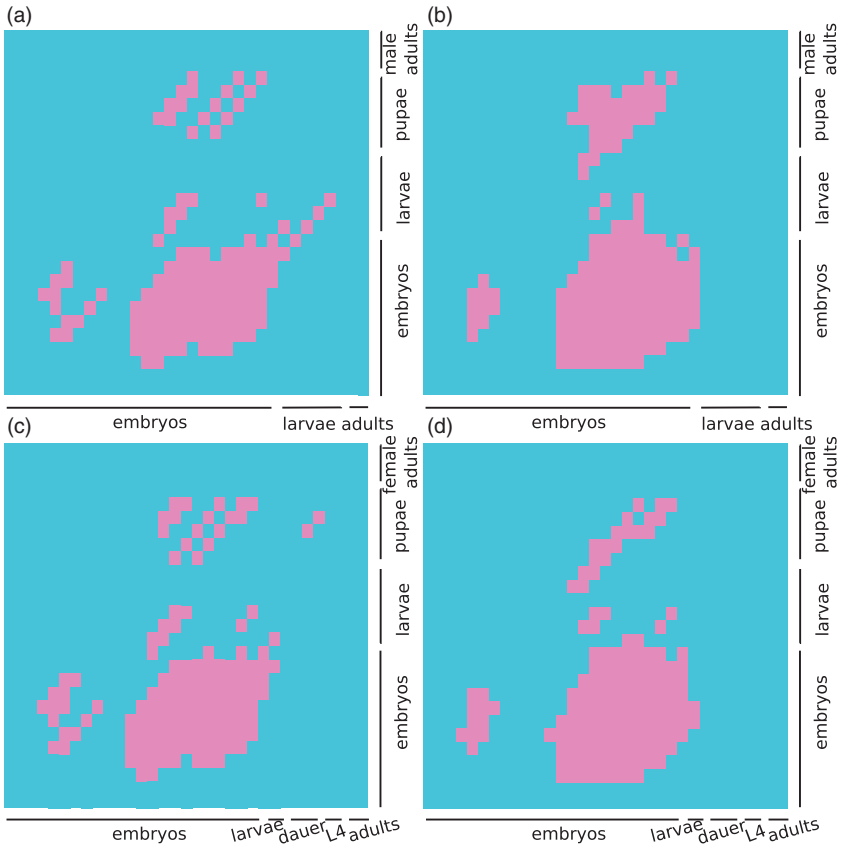

Fig. 4. Stage correspondence between fly and worm developmental stages using dataset combinations (a), (b) dm_male and ce_nd, and (c, d) dm_female and ce_da, limited to orthologous gene pairs (a, c) and extending to general gene pairs (b, d)

for different regions of k and d as the number of time points available grows.

As demonstrated in simulation, our method has a better power than LSA when correlated subsequences are broken into shorter segments separated by independent parts. On the other hand, on simple patterns with only one pair of correlated subsequences, the power of our method tends to decrease faster than LSA with increasing level of noise due to the latter using PC to measure association locally. It follows one natural extension of our method is to replace the indicators requiring exact rank profile matches with PC or Spearman's correlation. Such statistics would be more robust, but analysis of their asymptotic properties would be more challenging due to the overlaps between subsequences.

Applying the method to compare gene expression across developmental stages between *D. melanogaster* and *C. elegans*, we demonstrate in addition to being a measure of pairwise association, our method also directly provides mappings between the developmental timelines of the two species. The flexibility of our mappings allowing many-to-many correspondence between various time points is essential in cross-species analysis. The mapping results show two collinear patterns of correspondence, which is consistent with the findings in Li *et al.* (2014). More importantly, we are able to extend our analysis beyond orthologous gene pairs and replicate the key features of mappings using general gene pairs. As a measure of correlation, our method can be used to show orthologous gene pairs have stronger association statistically.

Last, the general applicability of our method is further illustrated in the computation of cross-correlated gene pairs between individuals with low and high LDLC levels. In this case, the lack of direct comparison between different individuals creates difficulties for correlation measures that focus on global trends; local measures including our method and the LSA are able to provide significant improvement. On the other hand, the PC outperforms the two local methods in the computation of within-sample correlations, suggesting the choice of appropriate correlation measure is context-dependent.

Funding

This work was supported by the Simons Institute for Theory of Computing, the National Science Foundation (1518001 to M.S.W., DMS-1160319 to H.H.), and the National Institutes of Health (U01-HG007031 to H.H., P50 GM115318 to M.M.). The generation of cell lines was supported in part by Diabetes Research Center (DRC) grant DK 0634591.

Conflict of Interest: none declared.

References

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

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

Basso,K. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.

Celniker,S.E. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. In *Ismb*. Vol. 8, pp. 93–103.

Daub,C.O. *et al.* (2004) Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Feng,J. *et al.* (2004) Time-frequency feature detection for time-course microarray data. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 128–132. ACM, New York.

Gao,Q. *et al.* (2012) Biclustering of linear patterns in gene expression data. *J. Comp. Biol.*, **19**, 619–631.

Goltsev,Y. and Papatsenko,D. (2009) Time warping of evolutionary distant temporal gene expression data based on noise suppression. *BMC Bioinformatics*, **10**, 1.

Hochreiter,S. *et al.* (2010) Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, **26**, 1520–1527.

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

Kato,M. *et al.* (2001) Lag analysis of genetic networks in the cell cycle of budding yeast. *Genome Informatics*, **12**, 266–267.

Kwon,A.T. *et al.* (2003) Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, **19**, 905–912.

Lazzeroni,L. and Owen,A. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, pp. 61–86.

Li,J.J. *et al.* (2014) Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.*, **24**, 1086–1101.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Pei,Y. *et al.* (2014) Identifying local co-regulation relationships in gene expression data. *J. Theor. Biol.*, **360**, 200–207.

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.

Ramoni,M.F. *et al.* (2002) Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, **99**, 9121–9126.

Roy,S. *et al.* (2014) Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics*, **15**, 1.

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

Simon,J.A. *et al.* (2006) Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *Am. J. Cardiol.*, **97**, 843–850.

Smith,A.A. *et al.* (2008) Similarity queries for temporal toxicogenomic expression profiles. *PLoS. Comput. Biol.*, **4**, e1000116.

Stegle,O. *et al.* (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500.

Steuer,R. *et al.* (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, S231–S240.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Wang,Y.R. *et al.* (2014) Gene coexpression measures in large heterogeneous samples using count statistics. *Proc. Natl. Acad. Sci. USA*, **111**, 16371–16376.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Xia,L.C. *et al.* (2013) Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics*, **29**, 230–237.

Yuan,Y. *et al.* (2011) Development and application of a modified dynamic time warping algorithm (DTW-S) to analyses of primate brain expression time series. *BMC Bioinformatics*, **12**, 347.