American Journal of Advanced Research, 2018, 2–1
December. 2018, pages 14-17
doi: 10.5281/zenodo.1419822
http://www.ibii-us.org/Journals/AJAR/
ISBN 2472-9264 (Online), 2472-9256 (Print)

Short Communication

An improved distance metric for clustering gene expression time-series data

Philip Heller* and Bharath Baiju

Department of Computer Science, San Jose State University, San Jose, CA 95192-0249.

*Email: philip.heller@sjsu.edu

Received on March 29, 2018; revised on September 14, 2018; published on September 16, 2018

Abstract

Gene expression in cells can fluctuate over time in response to internal or external stimuli. Time series expression studies can provide much information about how organisms function and respond to the environment. Due to economic constraints, such studies usually contain few time points and many genes; therefore, traditional time-series analysis techniques are not applicable.

Genes with similar expressions patterns generally share common function. Unsupervised clustering of genes on the basis of expression pattern can provide valuable insight functional relationships among genes. This insight is particularly valuable in the case of environmental bacteria, where the function of most genes of most species is unknown. Clustering requires a distance metric to quantify the pairwise differences among objects being clustered. In the case of gene expression, the most common metric is Pearson's Correlation Coefficient (PCC). Despite its popularity, PCC has a number of drawbacks: it does not match up with intuitive notions of distance, and it is insensitive to timepoint ordering. Consequently, clusters computed on the basis of PCC can contain dissimilar members and can lead to erroneous conclusions.

We propose a new metric, called ABLIM ("Area Between Linear Interpolations of Measurements"), which overcomes the shortcomings of PCC. ABLIM is visually intuitive, obeys triangle inequality, and is sensitive to timepoint ordering. Comparison of ABLIM and PCC clustering of gene expression data for the marine bacterium *Crocosphaera watsonii* demonstrates that ABLIM-based clusters more reliably reflect biological reality.

Keywords: Gene expression, time series, diel, clustering.

1 Introduction

24-hour cycling of gene expression has been observed in all three domains of life. This diel variation confers advantage by optimizing the interactions between external dark/light conditions and internal metabolic processes (Dodd et al., 2005). For example, in photosynthetic bacteria (cyanobacteria), photosystem proteins have half-lives of less than 12 hours (Yao et al., 2012), and their production generally begins a few hours before dawn; the organism is thus ready to harvest sunlight as soon as the sun rises. Production diminishes during the afternoon and ceases through most of the night. Diazotrophic cyanobacteria, which both photosynthesize and reduce atmospheric nitrogen (N_2), need to segregate nitrogenase enzymes from the oxygen produced by photosynthesis (Parker & Scutt, 1960; Bond, 1961; Fay, 1992), because oxygen is toxic to nitrogenase. Segregation is sometimes temporal, with nitrogenase component proteins produced 12 hours out of phase from photosystem proteins (Sherman et al., 1998).

The genomes of cyanobacteria, and especially of diazotrophic cyanobacteria, contain many other genes whose expressions also fluctuate on a daily cycle; the reason for the fluctuation is generally not known. Nevertheless, expression fluctuation can be a useful tool for studying cell function of these bacteria. Any tool that can shed light on cyanobacterial gene function is valuable, because for most cyanobacteria the function of at least 1/3 of the genome is unknown. Understanding of these microbes is important because they play a vital role in the so-called "carbon pump" – a biogeochemical process that removes greenhouse carbon from the atmosphere and exports it to the ocean floor (Siegenthaler & Sarmiento, 1993).

Bacterial expression patterns are typically measured with microarrays, which can quantify the expression of any gene represented on the array by a genetic probe. Probes are inexpensive, and it is possible to represent the entire genome (i.e. several thousand genes) of a bacterial species on a single microarray. However, processing of microarrays is expensive, so time-series expression studies generally have fewer than 10 time points. The combination of many genes and few time points raises analysis challenges. Established time-series analysis tools such as Fourier analysis, cosinors, periodograms, and wavelets require many more time points in order to model expression fluctuation. For example, Fourier Analysis takes advantage of the fact that any continuous repeating signal can be represented as an infinite sum of sines and cosines whose coefficients are functions of the original signal. For a discretely sampled input signal, the signal can be represented as a sum of n sines and cosines, where n is the number of time points in the experiment. When input signals are cheap and frequent, the resulting Fourier model is useful. For example, analyses of electrocardiogram or electroencephalogram studies might involve hundreds or thousands of data points (see for example Li et al., 1995; Polat & Gunes, 2006). By contrast, microarray gene expression studies typically have only 5-10 time points per day; therefore any individual gene's expression would be modeled by just 5-10 Fourier coefficients, which would not be enough to characterize thousands of gene expression profiles. Not only Fourier analysis but cosinors, periodograms, and wavelets have the same limitation. Evidently, novel approaches are needed.

A specific opportunity for development of novel analysis lies in unsupervised clustering of genes on the basis of similarity of expression profile. Genes with highly similar expression profiles generally have related function. Thus the function of a gene with unknown function can be predicted if the gene's profile is highly similar to that of a gene with known function. However, progress in this direction is hampered by the lack of a useful quantitative definition of expression profile similarity.

Computational tools that compare gene expression profiles have historically used Pearson's Correlation Coeffiecient (PCC) (Pearson, 1895) as a distance measure. However, this measure has a shortcoming that makes it inappropriate for the kind of time-series analysis considered here. The PCC formula depends on the expression levels of genes, but is independent of the times when those levels were measured. Thus, for example, in Figure 1 both graphs show fictitious expression levels of a pair of genes, measured at 7 time points. The blue and red expression levels are the same in both graphs; the only difference is the time points at which expressions were measured. In the graph at the left, the red and blue profiles appear to be similar, and might have related function; in the graph at the right, the profiles diverge over most of the experiment and suggest that the gene functions are unrelated. Surprisingly, the PCC distance between the upper blue and red profiles is the same as the PCC distance between the lower blue and red profiles.

A further problem with PCC is that it is not guaranteed to adhere to the triangle inequality rule for metric spaces (Fréchet, 1906), which requires that given any 3 entities A, B, and C, dAB may not be greater than dAC+dBC. Thus when PCC is used as a basis for unsupervised clustering, counterintuitive clusters can be produced, especially when using additive-distance tree-building algorithms such as Neighbor-Joining (Saitou & Nei, 1987) or Fitch-Margoliash (Fitch & Margoliash, 1967).

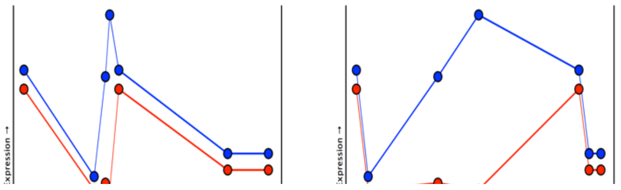

Fig. 1. Two expression graphs of two fictitious genes. The PCC distance between the blue and red profiles on the left is exactly equal to the PCC distance between the blue and red profiles on the right.

The shortcomings of the PCC metric can be observed in analysis of experimental data. Figure 2 shows the expression pattern of a gene (CwatDRAFT_3616, in blue) of the marine cyanobacterium Crocosphaera watsonii (Shi et al., 2010), along with the 10 most similar expression patterns (shown in red) as measured by PCC. Clearly in this case the PCC measure does not support our intuition of similarity.

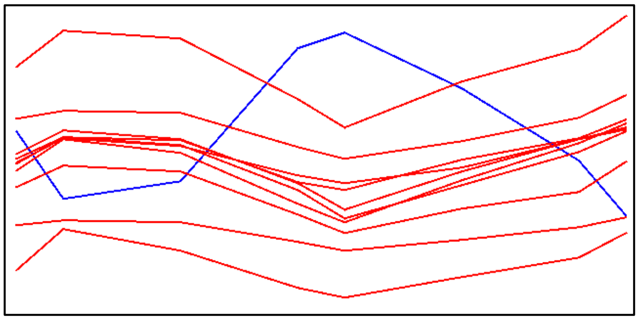

Fig. 2. Expression pattern of a gene (blue) of Crocosphaera watsonii, with expression patterns of the 10 most similar genes (red) according to PCC.

A similar shortcoming is observed when applying Fourier analysis to the same data set. Discrete Fourier Transform coefficients were computed for all expression profiles in the study. The difference between two profiles was defined straightforwardly as the Euclidean distance between the two vectors of coefficients. Profiles were retrieved for the 5 genes closest to nifH, a nitrogenase component; in addition to 3 related genes (nifD, nifK, and nifZ), these included coxB1 (an unrelated cytochrome c oxidase subunit), and opcA (an unrelated membrane protein).

We have developed a more intuitive distance measure, called ABLIM ("Area Between Linear Interpolations Metric"), for quantifying the difference between two expression profiles. Comparative analysis of diel gene expression in Crocosphaera watsonii using both PCC and ABLIM shows that ABLIM adheres to triangle inequality and is a more specific tool than PCC. To gauge ABLIM's efficacy, expression profiles of Crocosphaera watsonii were characterized using both PCC and ABLIM. ABLIM was demonstrated to be a more reliable distance measure.

2 Methods

The ABLIM distance measure is defined as follows: starting with measured expression patterns for 3 genes (red gene and blue gene, Figure 3A), add constants to the expressions of each gene so that both genes have the same mean expression (Figure 3B). Compute the linear interpolation of the expressions of each gene (Figure 3C). The ABLIM distance is the area between the interpolated patterns (Figure 3D, shaded area); this is easily computed as a sum of areas of triangles and parallelograms.

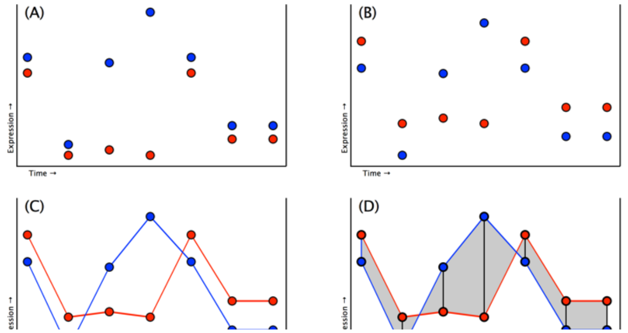

Fig.3. The ABLIM metric. (A) Original expression measurements. (B) each gene is normalized to a common mean. (C) linear interpolation. (D) ABLIM is the area (shaded) between the two curves.

Crocosphaera watsonii genes with diel expression were identified by filtering data from Shi et al. and retaining genes with at least 2x fluctuation over 24 hours. To estimate adherence to triangle inequality, 100,000 randomly selected sets of 3 diel genes were selected; distances within each set were computed by both PCC and ABLIM and checked for triangle inequality. Distances by both measures were then computed for all pairs of diel genes, pairs were binned by distance percentile, and bin populations were counted. Pairs were sorted by difference between PCC and ABLIM distance. Among genes where PCC similarity was high and ABLIM similarity was low, the 20 pairs with the most extreme differences between the 2 measures were manually inspected. Similarly, among genes where PCC similarity was low and ABLIM similarity was high, the 20 pairs with the most extreme differences between the 2 measures were also inspected.

3 Results

Of 100,000 randomly selected sets of 3 diel genes, 10,198 conformed to triangle inequality using PCC while all sets conformed using ABLIM.

Filtering produced 3,474 genes with diel expression; thus 12,065,202 pairs of genes were studied. Results of binning by distance percentile are shown in Figure 4.

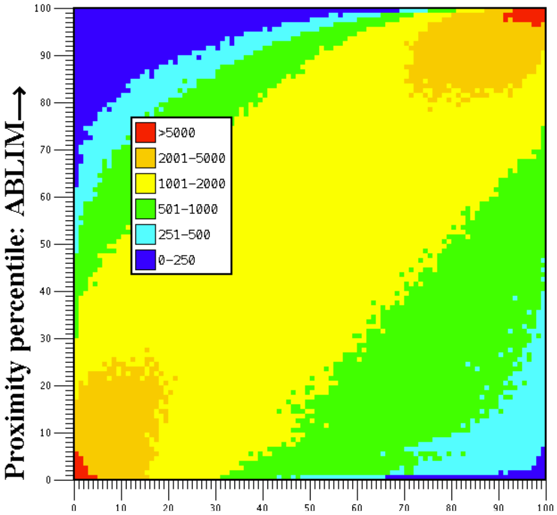

Fig. 4. Heat map of all pairwise distances between diel genes, binned by percentile.

Among the 20 gene pairs with the most extreme difference between high PCC similarity and low ABLIM similarity (bottom-right corner of Figure 4), the pattern shown in the top row of Figure 5 was typical: one gene (blue) exhibited a single high peak while the other (red) exhibited weak fluctuation with a low peak.

Among the 20 gene pairs with the most extreme difference between high ABLIM similarity and low PCC similarity (top-left corner of Figure 4), the pattern shown in the bottom row of Figure 5 was typical: both genes had low fluctuation.

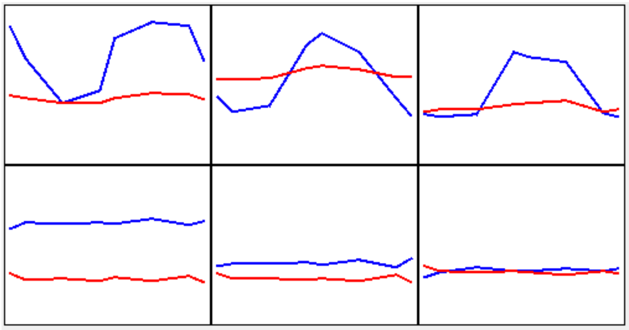

Fig.5. Gene pairs with extreme difference between PCC similarity and ABLIM similarity. Top row: High PCC similarity, low ABLIM similarity. Bottom row: high ABLIM similarity, low PCC similarity.

4 Discussion

The PCC measure conformed to triangle inequality in only 10% of 100,000 random sets of gene expressions. With the same data, the ABLIM measure conformed to triangle inequality in all cases. These results suggest that ABLIM is preferable for any distance-based clustering analysis of diel expression. For example, in a Neighbor-Joining or Fitch-Margoliash tree, the distance between any pair of nodes is supposed to reflect the actual distance between the physical entities represented by the nodes. Distances in the tree are computed by adding branch lengths; this addition is only meaningful if the metric that produced the distances conforms to triangle inequality. Thus trees built using PCC distances could be unreliable.

Figure 4 shows that when ABLIM similarity is high, PCC similarity is generally (but not always) high as well. Note that along the top edge of the figure where ABLIM similarity is at its maximum, the most populous (red) bins are at the top-right corner where PCC similarity is also at its maximum. Bins with maximum ABLIM similarity but low or moderate PCC similarity have low populations, as shown by the band of dark blue bins which extends horizontally well past the center of the figure. We conclude that high ABLIM similarity is generally supported by high PCC similarity. However, high PCC similarity is less often supported by high ABLIM similarity. Note that along the right edge of the figure where PCC similarity is at its maximum, the band of dark blue bins is short. Bins representing high PCC similarity but only low to moderate ABLIM similarity are significantly populated.

Given the significant contradictions between the two measures, it is important to determine which measure better conforms to intuition. The upper row of Figure 5 shows typical pairs of gene expressions where ABLIM similarity is low but PCC similarity is high; that is, PCC suggests a possible biological relationship, while ABLIM does not. In each pair, one gene has significant single-peak fluctuation while the other gene has nearly constant expression. Since closely related fluctuating genes tend to have coordinated fluctuation, the pairs shown in the upper row of Figure 5 are unlikely to be closely related. This cannot be verified in any of the cases in the upper row of Figure 5, where most of the genes have unknown function. However, Figure 6 shows a similar pair of expression profiles, where the functions of both genes are known, PCC similarity is high (92nd percentile), and ABLIM similarity is low (11th percentile). The blue line shows expression of hupL, which catalyzes hydrogen uptake; the red line shows expression of dnaA, which initiates DNA replication. These two functions are unrelated. Thus in this case ABLIM matches biological reality and PCC does not.

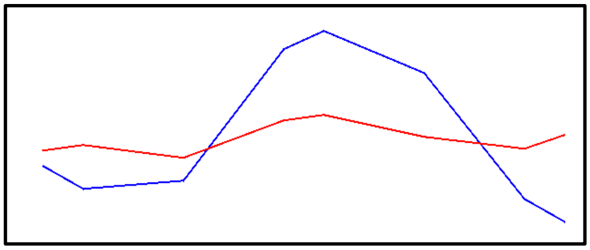

Fig.6. Expression profiles of hupL (blue) and dnaA (red). Gene functions are unrelated. PCC similarity is high, ABLIM similarity is low.

The lower row of Figure 5 shows typical pairs of gene expressions where PCC similarity is low but ABLIM similarity is high. In these cases gene expression is nearly constant, and PCC dissimilarity is due to small fluctuations that might be caused by noise. (Note that the pairs of profiles have mirror-image relationships, so that they are anti-correlated. A small noise-based change in any measurement could significantly change the correlation, which is reflected in the PCC.) Nearly constant-expression genes are not relevant to the intended use of the ABLIM metric, which is to provide a basis for reasoning about fluctuating genes.

To summarize these cases where PCC and ABLIM disagree about expression similarity: when PCC similarity is high and ABLIM similarity is low, ABLIM more accurately reflects biological reality; and opposite cases, when PCC similarity is low and ABLIM similarity is high, are not relevant to analysis requiring a similarity metric.

We conclude that ABLIM is the better tool, and has no disadvantages with respect to PCC. In future work we intend to use ABLIM to explore the *Crocosphaera watsonii* genome, and hope to develop techniques for inferring gene function based on expression similarity. It might also be possible, by using interpolation to simulate measured time points, to estimate the minimum number of microarray samples necessary for the application of traditional time series analysis.

Acknowledgements

The authors are grateful to Vincent Shen for his help with software analysis.

Conflict of Interest: none declared.

References

Bond,G. (1961).The oxygen relations of nitrogen fixation in root nodules. Z. Allg. Mikrobiol. 1:93-99.

Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., et al. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science*, 309(5734), 630–633. doi:10.1126/science.1115581.

Fay, P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol Rev*, 56(2), 340–373.

Fitch, W., & Margoliash. E. Construction of Phylogenetic Trees. *Science* 155, 279-284.

Fréchet, M. (1906). *Sur quelues points du calcul fonctionnel*. Doctoral thesis presented to la Faculté des Science de Paris.

Li, C., Zheng, C., & Tai, C. (1995). Detection of ECG Characteristic Points Using Wavelet Transforms. *IEEE T Bio-Med Eng*, 42(1).

Parker, C. A., & Scutt, P. B. (1960). The effect of oxygen on nitrogen fixation by Azotobacter. *Biochim Biophys Acta*, 38, 230–238.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *P R Soc London*, 58, 240-242.

Polat, K., & Gunes, S. (2007). Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Appl Math Comput*, 187, 1017-1026.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4), 406–425.

Sherman, L.A., Meunier, P., Colon-Lopez, M.S. (1998). Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth Res* (58): 25-42.

Shi, T., Ilikchyan, I., Rabouille, S., & Zehr, J. P. (2010). Genome-wide analysis of diel gene expression in the unicellular N2-fixing cyanobacterium Crocosphaera watsonii WH 8501. *ISME J*, 1–12. doi:10.1038/ismej.2009.148.

Siegenthaler, U., & Sarmiento, J.L. (1993). Atmospheric carbon dioxide and the ocean. *Nature*. 365, 119-125.

Yao, D. C. I., Brune, D. C., Vavilin, D., & Vermaas, W. F. J. (2012). Photosystem II component lifetimes in the cyanobacterium Synechocystis sp. strain PCC 6803: small Cab-like proteins stabilize biosynthesis intermediates and affect early steps in chlorophyll synthesis. *J Biol Chem*, 287(1), 682– 692. doi:10.1074/jbc.M111.320994.