Advance Access Publication Date: 29 November 2018

Prediction Interval Ranking Score: Identification of Invariant Expression from Time Series

Alexander M Crowell [1]*, Jennifer J. Loros [2] and Jay C Dunlap [1]*

[1] Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover NH, 03755, USA, and
[2] Department of Biochemistry and Cell Biology, Geisel School of Medicine at Dartmouth, Hanover NH, 03755, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Identification of constitutive reference genes is critical for analysis of gene expression. Large numbers of high throughput time series expression data are available, but current methods for identifying invariant expression are not tailored for time series. Identification of reference genes from these data sets can benefit from methods which incorporate the additional information they provide.
Results: Here we show that we can improve identification of invariant expression from time series by modelling the time component of the data. We implement the Prediction Interval Ranking Score (PIRS) software, which screens high throughput time series data and provides a ranked list of reference candidates. We expect that PIRS will improve the quality of gene expression analysis by allowing researchers to identify the best reference genes for their system from publicly available time series.
Availability: PIRS can be downloaded and installed with dependencies using 'pip install pirs' and Python code and documentation is available for download at https://github.com/aleccrowell/PIRS.
Contact: alexander.m.crowell@gmail.com, jay.c.dunlap@dartmouth.edu

Introduction

The identification of invariantly expressed reference genes underlies many studies in molecular biology. With the advent of widespread qRT-PCR studies, several tools were developed to identify constitutive reference genes from these experiments Vandesompele *et al*., 2002; Andersen *et al*., 2004; Pfaffl *et al*., 2004. More recently, several tools have been developed for identifying stably expressed reference genes from high throughput experiments Hruz *et al*., 2011; Carmona *et al*., 2017. Unfortunately, some reference genes thought to exhibit constitutive expression have later been found to be dynamically expressed in time series. GAPDH in particular was employed widely as a reference gene - and is still used in some cases - in spite of its circadian expression in many model organisms including mice Shinohara *et al*., 1998; Kosir *et al*., 2010; Pizarro *et al*., 2013. High quality time series expression data sets are available for a wide array of model organisms, yet reference genes are often still identified by analyzing small numbers of genes with tools not designed for time series Mughal *et al*., 2018; Jain *et al*., 2018. We seek to address this issue by leveraging the structure of time series experiments in order to better identify invariant expression.

PIRS first screens out clearly dynamic expression profiles using an ANOVA. The algorithm then looks for invariant expression both within and between time points for an expression profile by comparing mean expression to the prediction intervals of a linear regression. We report the results of simulations comparing the performance of PIRS to the more commonly used selection using the standard deviation/coefficient of variation. In these simulations, PIRS out performs methods which do not model the time component of the data. PIRS has previously been used to identify reference genes in *N. crassa* Hurley *et al*., 2015.

Methods

Algorithm Design

PIRS first performs an ANOVA on each expression profile to determine if any time points are differentially expressed and screens them out. This screening step both improves precision by removing clearly unwanted profiles from analysis and improves running time by reducing the number of profiles for which a score must be calculated. The algorithm then performs a linear regression of expression profiles y against the n time points x and calculates prediction intervals for these regressions. With prediction intervals:

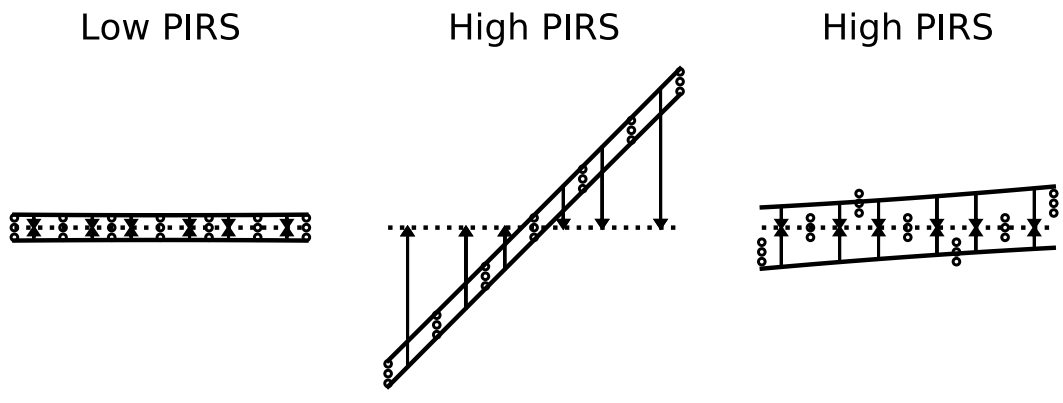
$$\hat{y} \pm t_{n-2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(1-n)s_x^2}}$$

where:

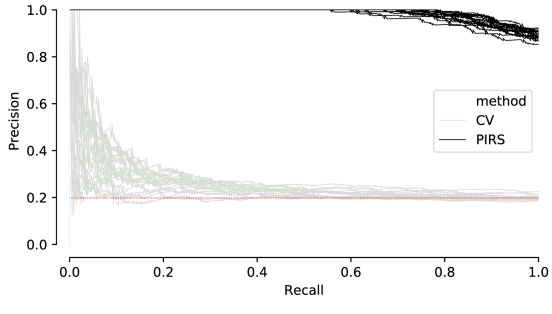
$$s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

and t_{n-2} is the $1 - \alpha/2$ ppf of the student's t-distribution with $n-2$ degrees of freedom.

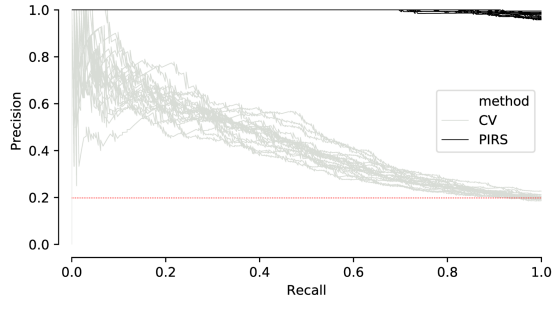
The absolute difference between the mean expression of the profile and the prediction intervals is then calculated for all time points. The final ranking score is derived by summing these differences and dividing by the mean expression. A truly constitutive expression profile would be expected to remain close to its overall mean across all observations and also to have relatively low variance around that mean. We would therefore expect the prediction intervals in this case to remain close to the overall mean for all observations. In contrast, we would expect a linear expression profile



(a) Example PIRS Calculations



(b) Precision Recall for Linear Trends



(c) Precision Recall for Circadian Trends

Fig. 1. PIRS: Description and Performance

(a) Example PIRS calculations. Circles indicate data points, dashed horizontal lines indicate mean of all data, solid lines show prediction intervals and arrows show the difference between prediction interval and mean. (b) PIRS performance on linear trends. Dark lines show precision/recall for PIRS, light lines show precision/recall for the standard Coefficient of variation method. Dashed horizontal line represents a random classifier. (c) Results for Circadian trends. Dark lines show precision/recall for PIRS, light lines show precision/recall for the standard Coefficient of variation method. Dashed horizontal line represents a random classifier

to have narrow prediction intervals which are far from the mean at most time points and a circadian expression profile to have prediction intervals centered on the mean but which are nonetheless far from it because they are broadly spaced (Fig 1 a).

Simulation Studies

In order to assess the effectiveness of PIRS at distinguishing constitutive from dynamic expression profiles, we generated simulated data sets and benchmarked PIRS against a standard deviation based method. We compared performance on both circadian and linear trends with gaussian noise added to all profiles. In both cases, 20% of profiles were truly invariant. In both tests, PIRS markedly outperformed the standard deviation based method (Fig 1 b and c).

Conclusion

When analysing time series data, the replicate and time point structure provide valuable information. PIRS provides improved results when screening time series data for invariant expression by incorporating this information and tailoring the analysis to the data. With PIRS, researchers can effectively identify reference genes from high throughput time series data sets in their model system. By identifying better reference genes, PIRS enables better down stream experiments.

Funding

Funding for AMC was provided by NIH grants R01GM083336 and 1R35GM118022 to JJL, 1R35GM118021 and 1U01EB022546 to JCD, and through an Albert J. Ryan Fellowship.

References

Cormona, Rosario *et al*. (2017) Automated identification of reference genes based on RNA-seq data, *BioMed Eng Online*, **16**:1, 65.

Andersen, Claus L *et al*. (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets, *Cancer Research*, **64**:15, 5245-5250.

Hruz, Tomas *et al*. (2011) RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization, *BMC Genomics*, **12**, 156.

Hurley, Jennifer M *et al*. (2015) A Tool Set for the Genome-Wide Analysis of *Neurospora crassa* by RT-PCR, *G3: Genes, Genomes, Genetics*, **5**:10, 2043-2049.

Jain, Nitin *et al*. (2018) Validation of house-keeping genes for normalization of gene expression data during diurnal/circadian studies in rice by RT-qPCR, *Nature Scientific Reports*, **8**, 3203.

Kosir, Rok *et al*. (2010) Determination of reference genes for circadian studies in different tissues and mouse strains, *BMC Mol Biol*, **11**, 60.

Mughal, Bilal B *et al*. (2018) Reference gene identification and validation for quantitative real-time PCR studies in developing Xenopus laevis, *Nature Scientific Reports*, **8**, 496.

Pfaffl, Michael W *et al*. (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper âŁ“ Excel-based tool using pair-wise correlations, *Biotechnology Letters*, **26**, 509-515.

Pizarro A*et al*. (2013) CircaDB: a database of mammalian circadian gene expression profiles, *Nucleic Acids Res.*, **41**, 1009âŁ“13.

Shinohara, Mari *et al*. (1998) Glyceraldehyde-3-phosphate Dehydrogenase Is Regulated on a Daily Basis by the Circadian Clock, *Journal of Biological Chemistry*, **273**, 446-452.

Vandesompele, Jo *et al*. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, *Genome Biology*, **3**:7.