CellPress

Time is Money: Designing Cost-Effective Time Series Experiments

Donna K. Slonim[1,2,*]
[1]Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA 02155, USA
[2]Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA
*Correspondence: slonim@cs.tufts.edu
http://dx.doi.org/10.1016/j.cels.2016.07.008

A new theoretical model helps to evaluate the tradeoffs between running technical replicates in high-throughput experiments and sampling at more time points.

Much has been written about experimental design for high-throughput gene expression experiments. Although technologies have changed over the years, many crucial design issues have persisted. A key question that often arises is how best to deploy limited resources to detect biologically meaningful changes in expression. In this issue, Sefer et al. (2016) address this question specifically for time series data.

Replication has long been known to be essential to identifying reproducible changes in expression data (Yang and Speed, 2002). However, exactly what "replication" means is a complex issue. Biological replicates typically represent different individuals or cellular samples under the same conditions, often embodied by individual patients or animals in each phenotypic group, while technical replicates are used to measure assay-induced variation with no biological import. For experiments comparing the expression of many genes in two or more categorical conditions, it has been shown that sampling a larger number of individuals is typically more powerful than running technical replicates or even, in the case of RNA-seq, than sequencing to greater depth (Churchill, 2002, Robles et al., 2012). The added power apparently derives from the fact that biological replicates also include variation from technical sources. Because adding individuals can capture both technical and biological variation at once, this can be the most cost-effective replication to perform, provided that there is no desire to separately quantify technical variability.

But for time series experiments, thought to account for about one-third of the expression datasets (Ernst and Bar-Joseph, 2006) in the Gene Expression Omnibus (GEO) database, there are special concerns. Sampling in this context could include replicates not just of technical processes or of individual organisms, but of different time points. In the most straightforward case, assuming a budget that can pay for n samples, one could assess expression at n equally spaced time points with a single sample at each. Sefer et al. call this a "dense" sampling strategy. Their alternative strategy, "replicate" sampling, describes choosing n/k time points but running k replicates at each.

The authors have developed a probabilistic model that allows them to evaluate the expected error in predicting gene expression at unmeasured time points using dense and replicate sampling strategies. The model makes specific but plausible assumptions about the underlying gene expression curves and level of variation. Temporal expression patterns are represented as piecewise-linear curves, i.e., curves approximated by adjoined straight lines. For example, Figure 1A shows two piecewise-linear curves representing two characteristic expression patterns in the yeast cell cycle (Tamayo et al., 1999): one that represents sinusoidal expression and another that suggests rapid upregulation followed by gradual decay. Although most gene expression patterns aren't actually piecewise-linear, the representation has proven a useful approximation in practice (Ernst et al., 2005).

The Sefer et al. model is designed to assess the detection of expression "jumps" of a fixed unit size, with expected technical variation parameterized as a function of that size. The authors then applied the model to predict expression at unsampled time points—first in simple step functions, then in more complex functions with multiple transitions, and finally in a real dataset. The results demonstrate that dense sampling is often better than replicate sampling provided that the technical noise isn't too high. Furthermore, dense sampling is almost never noticeably worse than replicate sampling, suggesting that there is little downside to making this choice.

Figure 1B illustrates the intuition behind this result. It shows the possible time points sampled across a single cycle of the patterns in Figure 1A for a fixed budget of 12 samples and different sampling strategies. Some replicate sampling choices leave out essential information about the expression patterns. For example, a replicate sampling strategy with four replicates at each of three time points might sample only at times 3, 7, and 11, missing both the degree of expression variation and the key difference in the responses characterized by these two patterns.

As a practical illustration of the model's applicability, Sefer et al. also analyzed data from an RNA-seq experiment (Rund et al., 2011) that measured temporal genome-wide expression patterns in the mosquito *Anopheles Gambiae*, with the goal of detecting circadian cycles in gene expression. Withholding some of the actual data allowed for accurate assessment of prediction. In practice, the dense strategy proved

more effective than replicate strategies at detecting circadian genes.

One of the most important contributions of this work is that it is accompanied by the release of the sampling strategy software created by the authors. The paper describes how, after running a small number of samples to assess variance, users may apply this software to assess tradeoffs when designing their own experiments.

This work represents the first such contribution to assessing this design tradeoff for time series experiments, but it is unlikely to be the last. The most common application of gene expression analysis is the detection of differential expression between conditions. There are many methods that do this explicitly in time series data (discussed in Storey et al., 2005). One interesting direction for future work would be to explicitly quantify how these sampling strategies affect the detection of significant differential expression across time series. It would also be valuable to determine whether the results favoring dense sampling hold for the optimal detection of sets of co-regulated genes.

Finally, the paper implies that this analysis holds for other sorts of high-throughput genomic data. However, relatively few other data types fit this particular model. Verification on some similar real datasets measuring, e.g., protein expression or ChIP-Seq counts, would be interesting. Determining just how broadly this model is applicable is an important open question.

Overall, Sefer et al. have demonstrated that, for many purposes, we can get more bang for our buck by sampling at additional time points rather than by running multiple replicates of the same time points. Evidence suggests that most experimentalists have already chosen to do this (Zinman et al., 2013), but the choice hasn't previously been justified. This work provides overdue confirmation that their decisions were right.

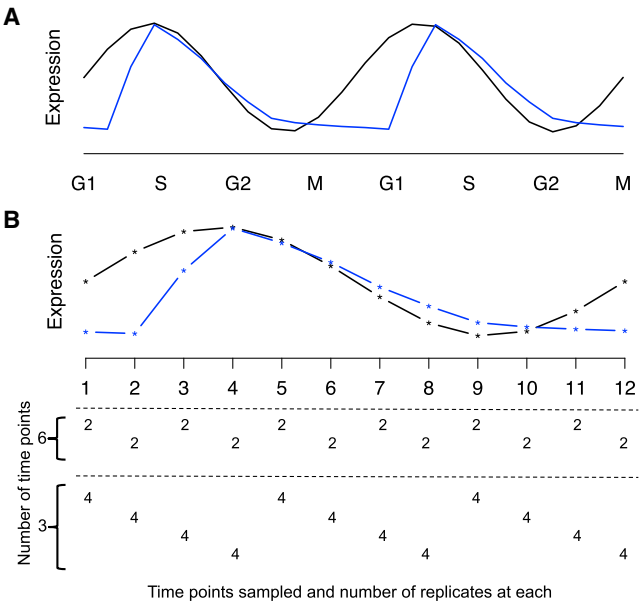

Figure 1. Sampling Strategies Illustrated by Hypothetical Cyclic Expression Patterns

(A) Piecewise-linear curves of two characteristic profiles of gene expression in the yeast cell cycle (Tamayo et al., 1999). The pattern shown in black represents sinusoidal expression over time, while the one in blue represents rapid upregulation followed by a gradual tapering.

(B) Sampling density influences the likelihood of detecting these patterns. Shown are 12 evenly spaced time points across a single cycle for the same expression patterns as in (A). A pure "dense" sampling strategy with a budget for 12 samples would have one sample at each time point, clearly characterizing the two expression patterns. Each row below the labeled time points shows the number of replicates and time points for a different "replicate" sampling strategy with the same budget. For example, the first row represents sampling six time points: 1, 3, 5, 7, 9, and 11, with two replicates at each. By omitting time points, some replicate strategies miss important differences between the two expression profiles.

REFERENCES

Churchill, G.A. (2002). Nat. Genet. *32* (*Suppl*), 490–495.

Ernst, J., and Bar-Joseph, Z. (2006). BMC Bioinformatics *7*, 191.

Ernst, J., Nau, G.J., and Bar-Joseph, Z. (2005). Bioinformatics *21* (*Suppl 1*), i159–i168.

Robles, J.A., Qureshi, S.E., Stephen, S.J., Wilson, S.R., Burden, C.J., and Taylor, J.M. (2012). BMC Genomics *13*, 484.

Rund, S.S., Hou, T.Y., Ward, S.M., Collins, F.H., and Duffield, G.E. (2011). Proc. Natl. Acad. Sci. USA *108*, E421–E430.

Sefer, E., Kleyman, M., and Bar-Joseph, Z. (2016). Cell Syst. *3*, this issue, 35–42.

Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., and Davis, R.W. (2005). Proc. Natl. Acad. Sci. USA *102*, 12837–12842.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). Proc. Natl. Acad. Sci. USA *96*, 2907–2912.

Yang, Y.H., and Speed, T. (2002). Nat. Rev. Genet. *3*, 579–588.

Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H., and Bar-Joseph, Z. (2013). Nat. Methods *10*, 925–926.