

An Active Learning Approach for Appropriate Sampling During Time-Series Expression Experiments

Rohit Singh^{1,2} Nathan Palmer^{1,3}

David Gifford^{1,4} Bonnie Berger^{1,5} Ziv Bar-Joseph^{6,7}

Keywords: microarray, time series, sampling, experiment design, active learning, spline

1 Introduction

Time-series expression experiments are becoming an increasingly popular method for studying a wide range of biological systems. One of the most important steps in designing such experiments is to define a sampling strategy. If the system is under-sampled, the results will not accurately represent the genes' activity over the duration of the experiment, and key features of the system's time-dependent response may be missed. On the other hand, over-sampling is expensive and time-consuming, and diverts resources that could otherwise be used for performing complementary studies.

To date, the determination of sampling rates for microarray experiments has relied mainly on the intuition of biologists. As such, sampling rates have differed amongst labs, even when studying the same biological phenomenon. Indeed, multiple experiments from the same lab, reported in the same paper, and directed at the same biological system have utilized different sampling rates. As mentioned above, these inconsistencies can lead to incomplete results, making it hard to compare data from related but independent studies.

2 Methods

A unique feature of microarray analysis, crucial to our work, is that biological samples may be frozen prior to hybridization, allowing researchers to extract the biological material following treatment at a very high rate, then make decisions about which samples to hybridize at a later time. We note that the expensive part of a microarray experiment is the hybridization step, rather than the act of extracting the sample. Thus, we can iteratively pick samples to hybridize, basing our decisions on data collected from previously sampled time-points.

We present the first online method for efficiently sampling during time-series microarray experiments. Beginning with an initial coarse sampling of the available time-points, our algorithm proceeds by first estimating a time-dependent expression profile in the form of a set of continuous functions that approximate the available data, as prescribed by Bar-Joseph *et al.*[1]. In order to quantify the uncertainty in these estimated profiles, we extend and adapt some recently-proposed statistical techniques for estimating error over spline-based smoothing functions [2]. By using local cross validation (LCV), we are able to focus in on localized signal variations, and appropriately consider the effect of sampling decisions, even

¹Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge MA 02139

²E-mail: rsingh@mit.edu

³E-mail: palmer@mit.edu

⁴E-mail: dkg@psrg.lcs.mit.edu

⁵Dept. of Mathematics, MIT, Cambridge MA 02139. E-mail: bab@mit.edu

⁶Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213

⁷Corresponding Author. E-mail: zivbj@cs.cmu.edu

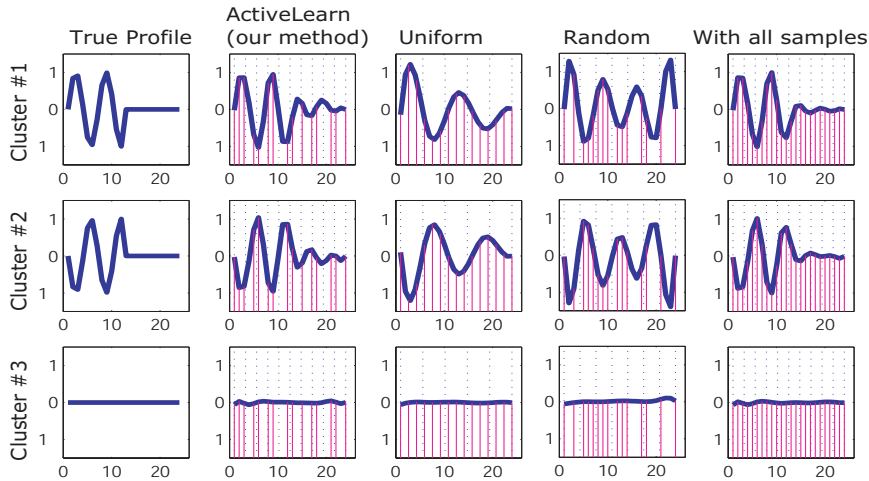


Figure 1: **Example Scenario:** we re-generate expression profiles for 150 genes (grouped into 3 clusters) by using various sampling strategies. The true profile, per cluster, is shown in the first column. A total of 24 sampling locations are potentially available. Columns 2-4 depict the performance of different sampling strategies when only 15 of these locations are actually chosen. The location of sampled time-points is indicated by the solid supporting bars (from X-axis to the curve). The fifth column shows results achieved by using all 24 samples. As can be seen, our sampling strategy (second column) performs much better than random sampling or uniform sampling, even with the same number of samples.

from non-uniform response data. We then use active learning to iteratively choose sample points, using the uncertainty in the quality of the currently estimated time-dependent profile as the objective function. One of our core contributions is the development of this efficiently-computable objective function for measuring the uncertainty in the estimated profiles. Our active learning approach allows us to identify the unsampled location at which the predicted observation will lead to the greatest reduction in overall uncertainty in the derived profile.

Because expression experiments profile thousands of genes at a time, it is infeasible to select new sample points on a per-gene basis; some genes may simply be irrelevant (e.g., non-cycling genes in cell cycle experiments). Instead, our algorithm evaluates expression profiles for clusters of co-expressed genes. We are in this way able to minimize the effect of noisy single-gene signals, and select new sample points that will contribute the most information to the expression profile as a whole.

3 Results

We have applied this method to both simulated and real biological data. Our algorithm performs well for both uniform and non-uniform response data. For biological data, we were able to reduce the number of time-points sampled by as much as 17% without significantly effecting the biological results.

References

- [1] Z. Bar-Joseph, G. Gerber, T.S. Jaakkola, D.K. Gifford, and I. Simon. Continuous representations of time series gene expression data. *Journal of Computational Biology*, 3-4:341–356, 2003.
- [2] D.J Cummins, T.G. Filloon, and Nychka D. Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *J Am Stat Assoc*, 96:453:233–246, 2001.