

# Effects of Sampling Rate and Time Duration for Noisy Time Series with Application to Microarray Experiment

Lingling An<sup>1</sup>, Fei Peng<sup>2,3</sup>, James Lu<sup>3</sup>

**Keywords:** microarray, time series, noise, experiment design, sampling, cell cycle

## 1 Introduction

Cell cycle is a basic biological phenomenon that the genetic contents replicate and cells grow, duplicate and split. Many critical biological problems and research questions including cancer are related to the cellular process. With the aid of microarray technology, there is an increasing interest in using the vast amount of microarray data to invest these problems. In particular, time-series expression experiments are widely used since biological systems and the related processes are time-varying. Different species may have different lengths of cell cycle. How to sample the cell cycles is a crucial issue [2, 3]. Over sampling may lead not only to high cost in experiment but also to the “fake” cell cycles because the microarray data are usually very noisy. On the other hand, under sampling may not reveal the correct cell cycles, i.e., the true activity of genes may be missed or represented bias. Determining sampling rate is quite discussed in signal processing but little research has been done on the noisy time series (i.e., signal with noise).

In this research, two problems are studied: (1) given a certain number of time points for a noisy signal with a certain period, what is an appropriate sampling rate? In other words, how many points are needed in one cycle? For example, if 64 time points are allowed to sample a signal, which way is better to accurately catch the cycle information, 32 cycles with 2 points per cycle, or 2 cycles with 32 points per cycle? (2) for a given noisy signal, how many time points in total do we need and what is the sampling rate, in order to gain the correct cycle information at a certain acceptable level? Throughout this research the frequency (i.e., reciprocal of period) of signals is used as a key feature since it is an indicator to relate to cell cycle activities.

## 2 Methods

Spectral analysis is employed to detect the frequency of signal. As gene expression data may contain great noise there may be more than one frequency found in a signal with a single frequency. The importance of frequencies detected in a signal are demonstrated by the spectrum, that is, a big spectrum may be corresponding to the true frequency while small spectrum may be the result of noise. When noise becomes great, the spectrum of frequency related to this noise is big too. So the problem turns to how to decide which frequency is meaningful, and which frequency is the result of noise. The data-driven method [1] is applied to statistically choose the meaningful frequency and disregard others.

A simulation study is performed to investigate the problems addressed above. For the first problem, consider a case of 64 time points to sample a signal which has a frequency of 1 Herz. we study all combinations of different sampling rate (2, 4, 8, 16, 32, 64) and different level of noise ([0, 0.1, , 1]). The noise is assumed normally distributed and the noise level is the standard deviation of the distribution. We define the power as: power=0 if the true frequency is not detected, otherwise power=1/number of found frequencies.

For the second problem, we use the similar parameter settings. Figure 1 shows the result of the first problem and Figure 2 is an example result of the second problem where the noise level is 1. In both figures the power displayed in Z-axis is the results of averaged power for 1000 signals.

## 3 Results and Discussions

The simulation results show that the noise level affect the power (Figure 1), which is not a surprise. Interestingly, for a fixed noise level in this graphics, the power for sampling rate of 2 is always the lowest while the change among others is not obvious. From Figure 2, it is found that the powers are quite close for some numbers of

<sup>1</sup>Department of Agricultural and Biosystems Engineering, University of Arizona, AZ, USA. E-mail: [anling@email.arizona.edu](mailto:anling@email.arizona.edu)

<sup>2</sup>Department of Computer Science, Shandong University, P. R. China. E-mail: [peng\\_intel@163.com](mailto:peng_intel@163.com)

<sup>3</sup>Department of Statistics, Purdue University, IN, USA. E-mail: [lu25@stat.purdue.edu](mailto:lu25@stat.purdue.edu)

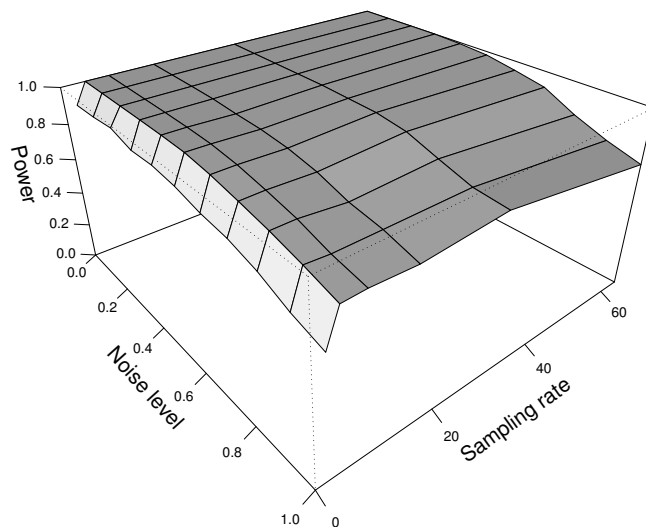


Figure 1: Power study of sampling rate for signals with 64 time points at different noise level.

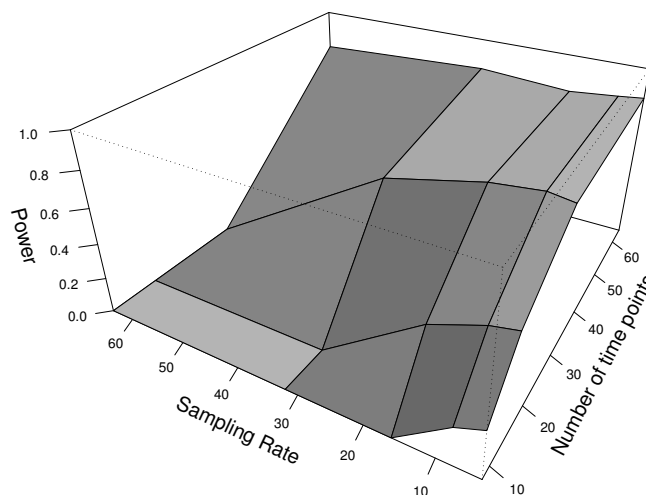


Figure 2: Power study of number of time points and sampling rate for signals with noise level 1.

time points, for example, the cases of 32 time points with sampling rate 8 and 64 points with sampling rate of 16. Similar results are obtained for signals at other noise levels (figures not shown).

According to the findings discussed above, some suggestions on experiment design can be made for researchers when they study the cell cycle expression profiles. Given a certain number of time points, i.e., certain number of gene chips, the sampling rate does not effect much except for the case of 2 points in one cycle. Another suggestion is that increasing the number of time points further cannot significantly increase the power when the number of time points is big enough. For example, increasing 32 points to 64 points does not increase power much but doubles the cost. So this research brings us some insights of experiment design when the research budget is tight.

## References

- [1] An, L. 2008. Dynamic clustering of time series gene expression. *Ph.D Dissertation, Purdue University, West Lafayette, IN, USA*
- [2] Bay, S. D. et al. 2003. Temporal aggregation bias and inference of causal regulatory networks. In: *Proceedings of the IJCAI workshop on Learning Graphical Models for Comp. Genomics.*
- [3] Bar-Joseph, Z. et al. 2003 Continuous representations of time series gene expression data. *Journal of Computational Biology*, 3-4:341356