Analyzing the Long Term Cellular Effects of an Environmental Toxin using Short Time Series Microarray Data

Darlene Olsen[1]
Karen Hinkle[2]
Norwich University[1,2], 158 Harmon Drive, Northfield, VT 05663

Abstract
Time course microarray experiments capture expression levels over time to examine the dynamic interaction of gene expression. Often in temporal microarray experiments, there are a limited number of time points taken resulting in short time series data. This is an overview of some techniques, such as STEM and maSigFun, used to analyze differentially expressed genes in short time series microarray data. In particular, this examintion is on the statistical models used to analyze temporal gene expression variation resulting when yeast (Saccharomyces cerevisiae) cells are exposed to 3-trifluoromethyl-4-nitrophenol (TFM), an environmental lampricide utilized frequently in the Great Lakes and Lake Champlain basins.

Key Words: Time series, microarray, gene expression

1. Introduction

Microarray experiments measure the expression levels of thousands of genes simultaneously to study the effect of conditions or treatments on gene expression. Time course microarray experiments capture expression levels for a gene over time (the temporal profile of a gene) to examine the dynamic interaction of gene expression. Often in temporal microarray experiments, a limited number of time points are taken resulting in short time series data. Conventional methods of time series analysis are only applicable to data sampled at many time points. Thus, several algorithms have been developed to analyze differentially expressed genes in short time series microarray data. The short time series miner (STEM) algorithm was applied to data that was collected to assess, at the cellular level, the environmental implications of applying a pesticide to the Lake Champlain basin.

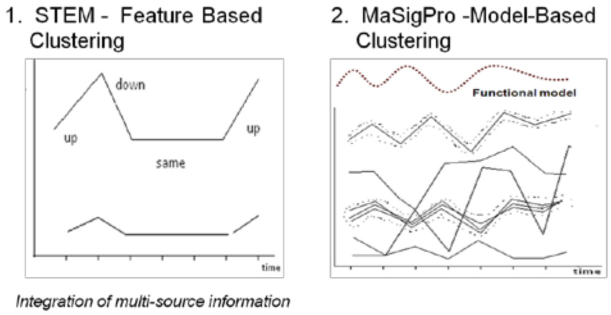
2. Background

2.1 Short time course microarray experiments analysis.

Microarrays measure gene expression levels to help reveal how multiple genes work together to respond to both static and dynamic conditions at the cellular level. Static experiments capture a single moment of gene expression. Time course experiments allow the temporal variations of gene expression to be examined. A vast amount of statistical work has been done to examine differential expression in static experiments. However, there are few statistical methods to examine temporal differential expression in short time series data [Chi, Y. *et al.* 2007].

The established methods for detecting differential expression in static experiments are not appropriate for microarray time series data. These methods fail to account for the inherent time dependence of the data [Storey, J. *et al.* 2005]. Conventional time series methods such as auto regression (AR), moving average (MA), Bayesian methods, or Fourier analysis modeling have been applied to microarray time series data. These methods were derived for long time series data and are not applicable to short time series [Kim, J. et al, 2007]. Due to the high costs of the arrays or limited biological samples, especially in clinical studies, the most common type of temporal expression data is short time series (typically fewer than ten sampled time points) [Wang, X., *et al.* 2008].

At the core of the analysis of expression data is the concept of similarity, and numerous methods have been examined to cluster time-series data. However, there does not seem to be a clear consensus on which algorithm to use that addresses all the challenges of short time series microarray data. Recent efforts to overcome the problems due to limited sampling include analysis using simplification strategies and integration of multi-source information [Wang, X. et al. 2008]. One method, short time-series expression miner (STEM), assigns temporal profiles to predefined clusters (Figure 1). Significant profiles are further examined using gene ontology (GO) enrichment analysis [Ernst, J. et al. 2005]. Another approach is a model-based clustering method that assumes expression profiles are clusters in the space of the functional forms (Figure 1) that represent them, and if the functional forms are similar, it is a result of the genes being involved in a similar cellular process (Androulakis et al., 2007). The maSigPro methodology is a model-based clustering algorithm that uses a two-regression step approach where the experimental groups are identified by dummy variables (Conesa et al., 2006).


1. STEM - Feature Based Clustering

2. MaSigPro -Model-Based Clustering

Integration of multi-source information

Figure 1: Schematic of clustering algorithm.

2.2 The short time series expression data.

The data was collected with the goal of understanding the molecular and cellular effects of chemical exposure of a pesticide on non-targeted organisms. For decades, 3-trifluoromethyl-4-nitrophenol, or TFM, (lampricide – pesticide) has been used in the Great Lakes and the Lake Champlain Basin to control the sea lamprey population. (Smith and Tibbles, 1980; Eshenroder et al., 1992). The goal of reducing sea lamprey numbers is to restore game fish populations that have shown a decline due to sea lamprey abundance. However, it is currently unclear whether other organisms in the Lake Champlain are being affected by TFM treatment, particularly those species that are endangered. (Gilderhus and Johnson, 1980; Matson, 1990; Nettles and Staats, 2001).

The time course microarray experiment exposed *Saccharomyces cerevisiae* (baker's yeast) with low doses of TFM (0.05mM) over time. Expression levels were measured at four time points 0 minutes, 60 minutes, 120 minutes and 240 minutes. There were two replicates at each time point and each replicate measures expression levels for approximately 5,800 genes.

3. Methods

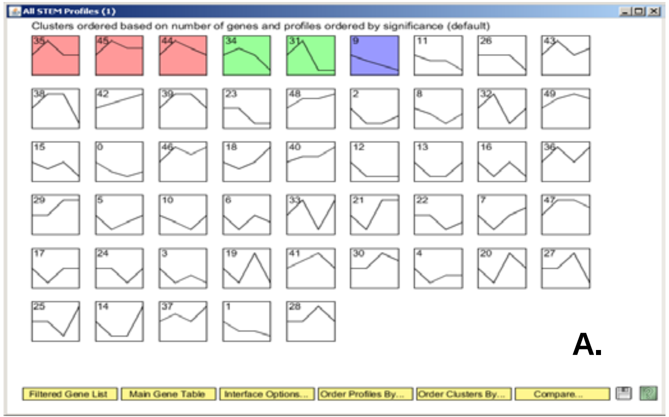
3.1 Methodologies for analyzing short time-series microarray data

Analysis of time course data has the potential to generate a wealth of information; however, there are many challenges to overcome. Methodologies need to account for multidimensionality, noisy data, non-uniform sampling points, and too few replicates (Bar-Joseph, 2004).

3.1.1 STEM

Short Time-Series Expression Miner (STEM) uses simplification strategies and integration of multi-source information (Ernst et al., 2005). STEM assigns temporal profiles to a set of representative, predefined profiles (inefficient for long time-series). Profiles are considered significant if they have a larger number of genes assigned than expected using a permutation test. Significant profiles are further examined using gene ontology (GO) enrichment analysis (Ernst et al., 2005).

Using STEM analysis (figure 2), DNA repair-related genes Ogg1, Pds5, and Rlf2, were clustered together in profile 35, while genes Pol30 and Cdc45 were clustered in profile 44. Both profiles were found to be significant and were grouped into a single cluster. Although Csm2 and Sae2 seem to visually share a similar profile with the genes clustered in profile 35, STEM did not identify those genes in any of the significant profiles. The fold change at 1 hour was lower for Csm2, and Sae2 had a slight upregulation between 2 and 4 hours. Quantitative PCR yielded results that were consistent with the short time course microarray in identifying an upregulation of DNA repair-related genes.



All STEM Profiles (1)

Clusters ordered based on number of genes and profiles ordered by significance (default)

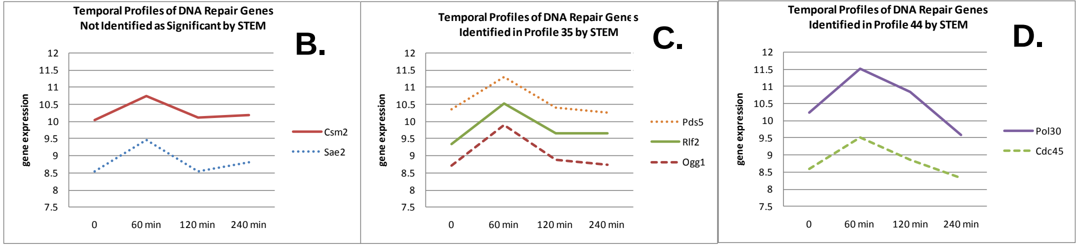
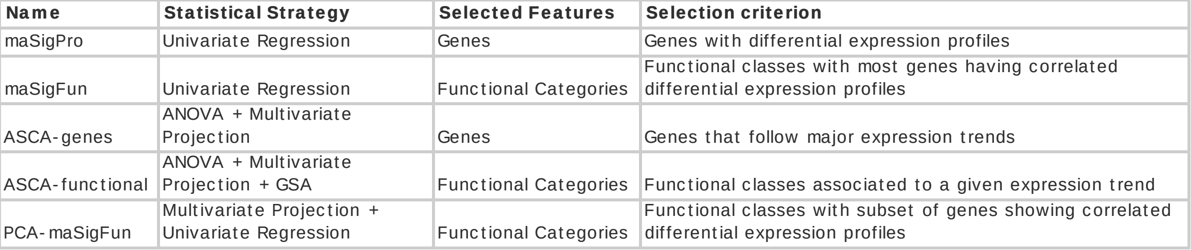


Figure 2: STEM profiles from short time-series TFM microarray study. **A.** The profile overview interface from the STEM software is shown. Profile ID number is shown at the top left corner of each box. Shaded profiles had a statistically significant number of genes assigned. Profiles of the same shade represent profiles grouped into a single cluster. **B-D.** Temporal profiles are shown for the seven DNA repair enzymes that were differentially regulated at one hour. **B.** Profiles of Csm2 and Sae2 are shown. **C.** Profiles for Ogg1, Pds5 and Rlf2 (STEM profile 35) are shown. **D.** Profiles for Pol30 and Cdc45 (STEM profile 44) are shown.

3.1.2 maSigPro

A general regression-based approach for the analysis of single or multiple microarray time series could be used to analyze these data. The methodology, named maSigPro (microarray Significant Profiles) is a two -step regression strategy that uses the model parameters to cluster genes (Conesa et al., 2006). Serial Expression Analysis (SEA) is a web site for the analysis of serial gene expression data and contains software for the maSigPro analysis (http://bioinfo.cipf.es/seawik). Figure 3 main characteristics of the SEA algorithms (taken directly from http://bioinfo.cipf.es/seawik). Currently, these algorithms are being used on the TFM data.



Name	Statistical Strategy	Selected Features	Selection criterion
maSigPro	Univariate Regression	Genes	Genes with differential expression profiles
maSigFun	Univariate Regression	Functional Categories	Functional classes with most genes having correlated differential expression profiles
ASCA-genes	ANOVA + Multivariate Projection	Genes	Genes that follow major expression trends
ASCA-functional	ANOVA + Multivariate Projection + GSA	Functional Categories	Functional classes associated to a given expression trend
PCA-maSigFun	Multivariate Projection + Univariate Regression	Functional Categories	Functional classes with subset of genes showing correlated differential expression profiles

Figure 3: Summary of the main characteristics of the SEA algorithms.

3. Conclusions

The results seem to vary depending on the algorithm used to analyze the data. Currently, maSigPro has not identified the DNA repair gene cluster identified by STEM. While the quantitative PCR yielded results that were consistent with the short time course microarray in identifying an upregulation of DNA repair-related genes, it is likely that significant temporal gene expression changes were missed between 0 and 1 hour, 1 and 4 hours, and 4 and 12 hours. Furthermore, in the microarray experiment, thousand of genes were profiled simultaneously with very few time points represented, making it difficult to determine significant biological changes versus patterns altered due to random chance (Bar-Joseph, 2004; Ernst et al., 2005). Since the study did not compare the TFM exposure samples against controls, it is difficult to conclude that changes in the regulatory process are due specifically to the TFM treatment. However, this preliminary study shows some of the limitations in attaining valuable information from short time series microarray experiments.

Acknowledgements

This work was partially supported by a Charles A. Dana Research Fellowship from Norwich University. This project was also supported by the Vermont Genetics Network through Grant Number 2P20RR016462 from the INBRE Program of the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

References

Androulakis, I., E. Yang, and R. Almon. 2007. Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. Annual Review of Biomedical Engineering, 9: 205-228

Bar-Joseph, Z. (2004) Analyzing time series gene expression data. Bioinformatics, 20(16): 2493–2503.

Chi, Y., Ibrahim, J., Bissahoyo, A., Threadgill , D. (2007) Bayesian Hierarchical Modeling for Time Course Microarray Experiments. *Biometric, Vol.* 63, pp. 496–504.

Conesa, A.; Nueda, M.J.; Ferrer, A. and Talón, M. (2006) maSigPro: a Method to Identify Significantly Differential Expression Profiles in Time-Course Microarray Experiments. *Bioinformatics*, 22 (9), 1096-1102.

Ernst, J., Bar-Joseph, Z. (2006) STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics, Vol 7:191.

Ernst, J., Nau, G., Bar-Joseph, Z. (2005) Clustering short time series gene expression data. Bioinformatics, Vol. 21, No. 1, pp. 159–168.

Eshenroder, R.L., Coble, D.W., Bruesewitz, R.E., Fratt, T.W., and Scheirer, J.W. (1992) Decline of lake trout in Lake Huron. *Trans of the Amer Fish Soc*, 121:548-554.

Gilderhus, P.A. and Johnson, B.G. (1980) Effects on Sea Lamprey (Petromyzon marinus) control in the Great Lakes on aquatic plants, invertebrates, and amphibians. *Canad J Fisher and Aquat Sc*i, 37: 1985-1905.

Kim, J. Kim,JH. (2007) Difference-based clustering of short time-course microarray data with replicates. *BMC Bioinformatics* 2007, Vol 8, 253.

Matson, T.O. (1990) Estimation of numbers for a riverine Necturus population before and after TFM lampricide exposure. *Kirtlandia*, 45:33-38.

Nettles, D.C. and Staats, N.R. (2001) A long-term program of sea lamprey control in Lake Champlain: Final Supplemental Environmental Impact Statement. *Fisheries Technical Committee Lake Champlain Fish and Wildlife Management Cooperative*, 153-159.

Smith, B.R. and Tibbles, J.J. (1980) Sea lamprey (*Petromyzon marinus*) in Lakes Huron, Michigan, and Superior: history of invasion and control, 1936-1978. *Can J Fish Aquat Sci*, 37:1780-1801.

Storey JD, Xiao W, T LJ, Tompkins RG, Davis RW (2005) Significance analysis of time course microarray experiments. *Proc Natl Acad Sci* , Vol. 102, pp. 12837–12842.

Wang, X., Wu, M., Li, Z., Chan, C. (2008) Short time-series microarray analysis: Methods and challenges. *BMC Systems Biology*, Vol. 2, 58