Short Time-Series Microarray Analysis: Methods and Challenges

Wang, Xuewei

BioMed Central

Wang, Xuewei, Ming Wu, Zheng Li, Christina Chan. "Short time-series microarray analysis: Methods and challenges" BMC Systems Biology 2:58. (2008)
http://hdl.handle.net/2144/3008

Boston University
**Abstract**

The detection and analysis of steady-state gene expression has become routine. Time-series microarrays are of growing interest to systems biologists for deciphering the dynamic nature and complex regulation of biosystems. Most temporal microarray data only contain a limited number of time points, giving rise to short-time-series data, which imposes challenges for traditional methods of extracting meaningful information. To obtain useful information from the wealth of short-time-series data requires addressing the problems that arise due to limited sampling. Current efforts have shown promise in improving the analysis of short time-series microarray data, although challenges remain. This commentary addresses recent advances in methods for short-time series analysis including simplification-based approaches and the integration of multi-source information. Nevertheless, further studies and development of computational methods are needed to provide practical solutions to fully exploit the potential of this data.

**Background**

Microarray technology has enabled the interrogation of gene expression data in a global and parallel fashion, and has become the most popular platform in the era of systems biology [1]. A majority of the microarray analysis thus far has focused on elucidating disease mechanisms [2]. More recently, with the rapid growth in research and development of biofuels [3], a new challenge of manipulating plant cell-wall biosynthesis has led to further applications of microarrays [3]. The detection and analysis of steady-state mRNA expression have become routine [4-7], with applications in many areas of biology (i.e., plants, yeast, insects, and mammals). Increasing efforts are focused on deciphering the multidimensional dynamic behaviours of complex biological systems, including complex regulation schemes, such as the crosstalk between multiple pathways [3,8,9], and interactions among more than 1000 genes in plant cell wall biogenesis, developmental biology, and human diseases [10-14]. Thus, time-series microarray data, and its analysis, are of growing interest to several research communities [15].

Time-series microarrays capture multiple expression profiles at discrete time points (i.e., minutes, hours, or days) of a continuous cellular process. These data can characterize the complex dynamics and regulation in the form of differential gene-expressions as a function of time. Numerous time-series microarray experiments have been performed to study such biological processes as the biological rhythms or circadian clock of *Arabidopsis*, flowering time, abiotic stress, disease progression, and drug responses [2,16-20]. Many of the methods of analyzing time-series data originated from various disciplines, such as signal processing, dynamic system theory, machine
learning and information theory, and have been applied to detect differentially expressed genes, identify expression patterns, and construct gene networks [15,21-23], nevertheless challenges remain.

A significant challenge in dealing with time-series data comes from the limited sampling or number of time points taken, giving rise to short time-series data. In the growing pool of temporal microarray datasets, a typical time-series record has fewer than ten time-points [24]. The most common type of temporal data available is short time-series data, which arises from the difficulty in obtaining samples for many time points, often times due to the high costs of the arrays or limited biological samples, especially in animal or clinical studies [25,26]. "Short" time-series could signify the time-scale or the number of discrete time-points. Typically, it refers to the latter, which more appropriately should be sparse time-series data.

Limited sampling accentuates the difficulties associated with static or standard time-series analyses. First, the problems arising due to high dimensionality accompanied by a small sample size, such as matrix singularity and model over-fitting [27], in analyzing static or long time-series microarray data, become more pronounced in the case of short time-series data. Second, the unavoidable noise has more influence on the analysis of short time-series than on long time-series data, enhancing the difficulty in distinguishing real from random patterns and increasing the potential of misleading analyses [28].

Improving short time-series analysis requires addressing the problems that arise due to limited sampling. Recent efforts by investigators to overcome the difficulties associated with limited sampling include decreasing the complexity of continuous time-series data based on simplification strategies [29,30] or enriching the information content of the data by incorporating multi-source information [31,32], see Figure 1 for a summary of possible options.

**Simplification strategies**

Simplification strategies reduce time-series data from continuous to discrete representations prior to analysis. These strategies usually transform the raw temporal profiles into a set of symbols [29,30,33] or nominal values [31,34] that are used to categorize qualitatively the gene expression data into different states or trends, that is, in terms of phases (early or late), magnitudes (high or low), or directions (up- or down-regulation). Based on this concept, a "quantization" method introduced by Di Camillo et al.[35], whereby the expression of a gene at a particular time-point is quantized (discretized) into three patterns of "states", representing under-expressed, not differentially expressed or over-expressed with respect to a baseline pre-defined by a hypothetical distribution. After such discretization, the Dynamic Bayesian Network algorithm performed better in terms of precision and recall in reconstructing the regulatory network from synthetic expression data generated from differential equations based on a series of defined rules of regulation. Similarly, Kim [33] developed a difference-based clustering algorithm (DIB-C) in which the profile of short time-series data was discretized to symbolic patterns, but according to the differences between adjacent time-points. These patterns or "trend" simplified the profile of a gene from numerical values to direction of change, that is, "I (Increase), D (Decrease) or N (No change)", and rate of change, that is, "V (conVex), A (concAve) or N (No change)". Inevitably information is lost through this simplification. Even so, such conceptual discretization helped achieve more interpretable and biologically meaningful clusters [33].

Simplifications methods have a side benefit in reducing the noise in the original data to some degree when decreasing the dimension of the time-series data, thus making the subsequent analysis more robust to noise. This was demonstrated by Sacchi et al. [30] with their adaptation of the Temporal Abstractions (TA)-clustering method from the field of artificial intelligence to gene expression analysis. Here, the temporal expression profiles were described in terms of trends of "Increasing", "Decreasing", or "Steady". A reduced rate of misclassification in computational experiments was observed for simulated data using TA-clustering with pre-defined patterns and noise than with the clustering approach without such simplification strategies, particularly when the noise level was high [30].

A key challenge with simplification strategies is how to pre-define these a priori representative temporal trends or patterns of gene expression in the discretization step. Defining these patterns have largely depended on the expertise of the researchers, for example, Gerber et al defined six temporal expressions trends in terms of phase (early, middle and late) and direction (increase and decrease) [31], similarly, Wu et al. proposed 27 possible temporal patterns to group gene expression data for CD8 T cell differentiation [34]. However, this may introduce bias in the patterns that are pre-defined and, in turn, the analysis and results obtained. Data-driven approaches could extract potentially novel gene expression patterns in an objective and reasonably unbiased fashion [36]. Thus, developing methods to automatically define temporal trends could alleviate this limitation or bias. Ernst et al. proposed a procedure to generate potential trends which describe the directions and magnitudes of the expression changes with respect to time [24,28]. Attempts at automatic abstraction of temporal features have met with some success in providing easily interpretable clusters,
examples include the temporal abstraction-based method that defines trends (i.e., Increasing, Decreasing and Steady) over subintervals [30], and the difference-based method that uses the first and second order differences in expression values to detect the direction and rate of change of the temporal expression [33]. Although simplification strategies make the raw expression profiles coarse-grained, which could somewhat ameliorate the noise in the data, inevitably the simplification leads to loss of information, which may exacerbate the situation of limited sampling.

In particular, some important patterns may be lost when the raw expression profiles are oversimplified, for example, simplifications that consider only monotonously expressing genes [31] may not capture some of the complex temporal patterns, such as oscillatory gene expression profiles [37].

Incorporating multi-source information
Incorporating multi-source information, including prior knowledge (i.e., pathway information) [38,39], multi-
scale or different levels of information [40-42], or additional time-series datasets from other sources [31,32], is another approach to address the limited sampling and to improve the computational analysis and interpretation of short time-series microarray data.

Different types of prior knowledge have been used to improve the computational analysis of short time-series data. They include applying a prior noise distribution to the expression data [43]. For example, by incorporating a prior noise-distribution to improve the parameter estimation in the commonly used CAGED model (Cluster Analysis of Gene Expression Dynamic), Wang et al. achieved more functional and meaningful clusters, as validated by Gene Ontology [43]. This approach was advanced further by Wang et al. [44] to a stochastic dynamic model where the gene expression profile is modelled with the addition of noisy "measurements". The authors try to explicitly separate the real pattern of expression from the Gaussian noise imposed on the expression data. Based on such a model, they applied Expectation Maximization (EM) algorithm to estimate both the parameters for the noise model and the actual values of the expression levels, and efficiently reconstructed the gene regulatory network. Thus defining a prior noise-distribution in analyzing time series microarrays is both biologically relevant and computationally efficacious especially when the time series is too short to satisfy the requirements of traditional multivariate methods for parameter estimation [44].

In addition, pre-defined gene sets involving specific pathways or functional categories have focused on pattern changes of sets of genes rather than individual genes and helped to enhance our understanding of cellular processes [38,39]. Similarly, incorporating multi-level biological information, such as metabolite data or prior knowledge about the genes and pathways, has improved interpretation of the data. For example, metabolic data [40,41] and pathway information [40,42] have been integrated with short time-series gene expression data to identify liver toxicity pathways in HepG2 cells. Likewise, protein-DNA interaction data and promoter motif information have been integrated with short time series data to reconstruct the dynamic gene regulatory network of Saccharomyces cerevisiae response to stress [45], and to identify targets of known transcription factors in cold acclimation of Arabidopsis thaliana [46], respectively. Furthermore, metabolic profiles have been integrated with short time-series gene expression data to characterize the dynamics of metabolic changes during oxidative stress [47], the effect of elevated CO₂ on the physiology of A. thaliana [48], and to reconstruct the temporal sequence of events during bud development [49]. Similarly, integrating multiple time-series datasets has become increasingly popular with the growing pool of publicly available datasets [50]. Combining multiple time-series datasets has been shown to improve the confidence of the gene regulatory relationships that are inferred [51], as well as identify regulatory relationships [32] and functional gene clusters [31] under different treatment conditions.

A key challenge with integrating different datasets is the heterogeneity of the data, that is, each set may have a unique set of sampling rates, time-scales, cell types, and sample populations, as well as varying measurement noise levels, etc. The heterogeneity across the datasets increases the difficulty in extracting meaningful results. To maximize the usefulness and minimize the heterogeneity of the publicly available data, stricter standardization methods should be defined and imposed on procedures such as data collection and pre-processing. Indeed, standards such as MIAME (Minimum information about a microarray experiment), MIAPE (Minimum information about a proteomics experiment), MSI (Metabolomics standards initiative), MIMIx (Minimum information required for reporting a molecular interaction experiment) have been proposed and implemented for presenting and exchanging gene expression [52], proteomics [53], metabolomics [54] and interaction data [55], respectively. Thus far, standardizing gene expression data is the most mature and hence, most successful compared to the standardization of the other data types. Therefore, integrating gene expression data from various sources is now readily achievable with public databases, such as GEO [56] and ArrayExpress [57], where the quality of the data is controlled with the MIAME score.

Conclusion
In summary, analysis of short time-series microarrays is still at an early stage. Most studies using short time-series data have applied methods that had been developed for static or long time-series microarray data, and which tend to perform poorly with limited temporal sampling. Current efforts, including simplification approaches and the integration of multi-source information, have shed promising light on improving the analysis of short time-series microarray data.

Future studies could combine both of these strategies to simultaneously decrease the complexity of continuous time-series representations, yet minimize the information loss with the simplification-based approaches by increasing the information content of the data. Gene-module-level analysis could be a potential solution, in which the concept of modularity not only plays a central role in incorporating multi-source biological information, but also reflect a simplification strategy focusing on groups of genes rather than individual ones. Gene-module-level analysis could efficiently combine both strategies.
A recent study by Hirose et al [58] used a statistical inference method to reconstruct a module-level gene network based on time-series data, rather than networks of individual genes. They concentrated on groups of genes and the correlations between them, thus the transcription modules extracted could be building blocks of the regulatory networks. Such module-based network construction overcomes, in part, the problem of limited sampling. The modules in the study are calculated by a vector regressive approach based on the state space model, which essentially simplifies the data by including only the significant temporal relationships between the modules. Unfortunately, their modules are defined based on statistical criteria and thus are limited in their biological significance. The integration of multi-source biological information to identify modules from short-time series microarray data should enhance understanding and interpretation of biological systems and disease processes.

Thus far, the predominant focus has still been on lower levels of analyses, such as detecting differently expressed genes or clustering genes with similar temporal profiles, whereas few higher levels of analysis, i.e. network construction, have been reported. With the rapid growth in availability of short time-series data, more theoretical and technical studies are urgently needed to provide practical solutions to exploit fully the potential of this wealth of data.

Acknowledgements
We thank Professor Neil T. Wright for providing critical comments on the content, and the editors for their valuable comments and suggestions in improving the paper. C.C is supported in part by the National Institute of Health (1R01GM079688-01), National Science Foundation (BES 0425821), and the MSU Foundation on the Center for Systems Biology.

References
1. Panda S, Sato TK, Hampton GM, Hogensjeb JB: An array of insights: application of DNA chip technology in the study of cell biology. Trends in cell biology 2003, 13(3):151-156.
2. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, Xiao W, Laudanski K, Brownstein BH, Elson CM, Hayden DL, Hernond DN, Lowry SF, Maier RV, Schoenfeld DA, Moldawer LL, Davis RW, Tompkins RG, Baker HV, Bankey P, Billiar T, Brownstein BH, Calvano SE, Camp D, Chaudy I, Cobb JP, Davis RW, Elson CM, Freeman B, Gibran N, Harbrecht B, Hayden DL, Heagy W, Heinbach D, Hernond DN, Horton J, Hunt J, Laudanski K, Lederer J, Lowry SF, Maier RV, Mannick J, McKinley B, Miller-Graziano C, Mindrinos MN, Minei J, Moldawer LL, Moore E, Moore F, Munford R, Nathens A, O'Keefe G, Purdie G, Rahme L, Remick D, Salmi M, Schonfeld DA, Shapiro M, Silver G, Smith R, Stepasopanoulis G, Stormo G, Tompkins RG, Toner M, Warren S, West M, Wolls S, Xiao W, Young Y: Application of genome-wide expression analysis to human health and disease. Proceedings of the National Academy of Sciences of the United States of America 2005, 102(13):4801-4806.
3. US Department of Energy, Office of Sciences: Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda. 2006.
4. Salunkhe P, Topfer T, Buer J, Tummler B: Genome-wide transcriptional profiling of the steady-state response of Pseudomonas aeruginosa to hydrogen peroxide. Journal of bacteriology 2005, 187(8):2565-2572.
5. Rosso D, Ivanov AG, Fu A, Geisler-Lee J, Hendrickson L, Geisler M, Stewart G, Krol M, Hurry V, Rodermel SR, Maxwell DP, Huner NP: IMPACTJANS does not act as a stress-induced safety valve in the protection of the photosynthetic apparatus of Arabidopsis during steady-state photosynthesis. Plant physiology 2006, 142(2):574-585.
6. Rawool SB, Venkatesh KV: Steady state model to approach gene regulatory networks—simulation of microarray experiments. Bio Systems 2007, 90(3):636-655.
7. Kocabas AM, Crosby J, Ross PJ, Otu HH, Beyhan Z, Can H, Tam WL, Rosa GJ, Halgren RG, Lim B, Fernandez E, Cicelli JB: The transcriptome of human oocytes. Proc Natl Acad Sci U S A 2006, 103(18):6427-6432.
8. Laule O, Furlholz A, Chang HS, Zhu T, Wang X, Heifetz PB, Grussew W, Lange M: Crosstalk between cytokolic and plastidial pathways of isoprenoid biosynthesis in Arabidopsis thaliana. Proc Natl Acad Sci U S A 2003, 100(11):6666-6671.
9. Andresen JR, Royce TE, Shomer A, Masquera JM, Demichelis F, Hofer MD, Mertz KD, Gerstein M, Rubin MA: Integrative Microarray analysis of pathways dysregulated in metastatic prostate cancer. Cancer Res 2007, 67(21):10296-10303.
10. Yong WD, Link B, O'Malley R, Tewari J, Hunter CT, Lu CA, Li XM, Bliekker AB, Koch KE, McCann MC, McCary DR, Patterson SE, Reiter WD, Staiger C, Thomas SR, Vermerris W, Carpta NC: Genomics of plant cell wall biogenesis. Planta 2005, 221(6):747-751.
11. Carpita N, Tierney M, Campbell M: Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. Plant Mol Biol 2001, 47(1-2):1-5.
12. Dozmorov MG, Kyrky KD, Saban R, Shankar N, Baghdyan AS, Centola MB, Hurst RE: Systems biology approach for mapping the response of human urethelial cells to infection by Enterococcus faecalis. BMC bioinformatics 2007, 8 Suppl 7:52.
13. Hooper SD, Boue S, Krause R, Jensen LJ, Mason CE, Ghanim M, White KP, Furlong EE, Bork P: Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. Mol Biol Evol 2007, 24(5):737-747.
14. Baugh LR, Hill AA, Sionim DK, Brown EL, Hunter CP: Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development (Cambridge, England) 2003, 130(5):899-900.
15. Androulakis IP, Yang E, Almon RR: Analysis of time-series gene expression data: Methods, challenges, and opportunities. Annual Review of Biomedical Engineering 2007, 9:205-228.
16. Hsu KL, Pilibello KT, Mahal LK: Analyzing the dynamic bacterial glycome with a lectin microarray approach. Nature chemical biology 2006, 2(3):153-157.
17. McAdams HH, Shapiro L: A bacterial cell-cycle regulatory network operating in time and space. Science 2003, 301(5641):1874-1877.
18. Lan H, Carson P, Provat NJ, Bonner AJ: Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. BMC bioinformatics 2007, 8:358.
19. Welch SM, Roe JL, Dong ZS: A genetic neural network model of flowering time control in Arabidopsis thaliana. Agron J 2003, 95(1):71-81.
20. Locke JC, Millar AJ, Turner MS: Modelling genetic networks with noisy and varied experimental data: the circadian clock in Arabidopsis thaliana. Journal of theoretical biology 2005, 234(3):383-393.
21. Bar-Joseph Z: Analyzing time series gene expression data. Bioinformatics (Oxford, England) 2004, 20(16):2493-2503.
22. Opgen-Rhein R, Strimmer K: Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC bioinformatics 2007, 8 Suppl 2:14.
23. Opgen-Rhein R, Strimmer K: From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol 2007, 1:31.
24. Ernst J, Bar-Joseph Z: STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics 2006, 7:191.
25. Ding M, Cui SY, Li CJ, Jothy S, Haase V, Steer BM, Marsden PA, Pippin J, Shankland S, Rastaldi MP, Cohen CD, Kretzler M, Quaggin SE: Loss of the tumor suppressor Vhl leads to upregulation of Cxcr4
31. Gerber GK, Dowell RD, Jaakkola TS, Gifford DK: 
32. Karpuj MV, Becher MW, Springer JE 
33. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z: 
34. Wang Z, Yang F, Ho DW, Swift S, Tucker A, Liu X: 
35. Chawade A, Brautigam M, Lindolff A, Olsson O, Olsson B: 
36. Baxter CJ, Redestig H, Schauer N, Repilser B, Patil KR, Nielsen J, Selbig J, Liu J, Fernie AR, Sweetlove LJ: 
37. H. Kanani, B. Dutta, J. Quackenbush, Klpa M: Time-Series Integrated Metabolomic and Transcriptional Profiling Analyses. 
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gilchrist M, Xiang K, Liu J, Borrello A, Petersen M, Schuler-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. 
39. Segal E, Friedman N, Koller D, Regev A: 
40. Li Z, Srivastava S, Findlan R, Chan C: Using Dynamic Gene Module Map Analysis To Identify Targets That Modulate Free Fatty Acid Induced Cytotoxicity. Biotechnology Progress 2008, 24(1):29-37. 
41. Li Z, Srivastava S, Findlan R, Chan C: Developing Dynamic Gene Module Map Analysis To Identify Targets That Modulate Free Fatty Acid Induced Cytotoxicity. Biotechnology Progress 2008, 24(1):29-37. 
42. Wang L, Ramani M, Sebastiani P: Clustering short gene expression profiles. Lect Notes Comput Sc 2006, 3909:60-68. 
43. Wang L, Yang F, Ho DW, Swift S, Tucker A, Liu X: Stochastic dynamic modeling of short gene expression time-series data. IEEE transactions on nanobioscience 2008, 7(1):44-55. 
44. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z: Reconstructing dynamic regulatory maps. Mol Syst Biol 2007, 3:74. 
45. Swan M, Vaidyanathan S, Sargent D, Houghton J, King M, Russell R, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR, Hermjakob H: The minimum information required for reporting a molecular interaction (MIMIx). Bioinformatics 2007, 23(8):894-898. 
46. Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo/]

ArrayExpress [http://www.ebi.ac.uk/arrayexpress/]

Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, Print C, Miyano S: Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics 2008, 24(7):932-942.

BMC Systems Biology 2008, 2:58

http://www.biomedcentral.com/1752-0509/2/58

Publish with BioMed Central and every scientist can read your work free of charge.

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:
- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp