A general review on time-series gene expression data

YIDA ZHANG and LE SHU, University of California, Los Angeles

Biological processes are often dynamic. To capture the dynamics of these biological processes, we need to monitor different time points during the processes. Time-series data is the data type which contains information of biological processes at different time points. Based on these information, time-series data can be used to investigate the whole process in a biological activity, to infer the rate of expression change, the order of changes and possible causal relationships. In this review, we will discuss basic experimental background of time-series data. Besides, we will also discuss the general procedure of analyzing time-series data in a computational way.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*; I.5.1 [**Pattern Recognition**]: Models—*Neural Nets*

General Terms: Human Factors

1. INTRODUCTION

As the dynamic property of biological processes, time-series data are very important for us if we want to understand and model complex biological processes. Although there are some types of genomic data also including information over time, time-series gene expression data is the most abundant and available one compared with other data types. Time-series gene expression data can be used to gain a wide range of insights. For example, it can be used to characterize the relationship between different genes, their regulation and coordination and also pathogenesis of complex diseases. Consequently, time-series gene expression data has been widely used and become an very important tool to investigate biological processes such as responses to external stimuli, disease progression and cell cycle. Recently, the emergence of methods for measuring gene expression such as high-throughput RNA sequencing (RNA-seq) and the increased focus on clinical applications make time-series gene expression studies more powerful and feasible. Moreover, the amount of time-series expression data in public expression databases has grown exponentially over the past few years [Barrett et al. 2013]. Besides, the increase in sequence capacity, which has been used primarily to generate static datasets, makes time-series expression date even more attractive as a powerful complementary for the understanding of dynamic systems.

However, although time-series data has many benefits, it also raises some experimental and computational challenges. In this review, we will go over these challenges and present the basic experimental considerations and computational methods that have been developed for analyzing time-series gene expression data. Of note, although this review is mainly focused on gene expression data derived from microarrays, most of the points discussed here can be applied to sequence-based gene expression data.

2. EXPERIMENTAL DESIGN

2.1 Advantage of time-series experiment

Before designing a time-series experiment, the question we should ask is what is the advantage of time-series experiment? Or to say, what are the benefits of multiple measurements rather than single and static measurements? First, time-series gene expression experiments can capture information about genes with transient expression changes. This can be applied for all types of biological processes. For example, in perturbation-response experiments, different sets of genes respond with different kinetics.

Consequently, the entire response can be captured only by monitoring the process at multiple time points [Gasch et al. 2000]. Second, time-series experiments can provide a view of the sequence of events that take place. This is important both in order to understand the type of processes that are activated at each stage and also for inferring causality. Finally, time-series experiment allows the study of the kinetics and the temporal pattern of a response. This is very important for understanding the dynamics of biological processes.

2.2 Biological conditions suitable for time-series experiment

Most genomic methods are applied to populations of cells rather than single cells. Technological advances have already allowed the measurement of the transcriptome of single cells, however, sampling a cell at multiple time points is still a problem. Consequently, time-series experiments are suitable to cases in which cell populations are fairly uniform. The following processes produce such populations: response to external signals [Gasch et al. 2000]; developmental processes with a clear starting point [Gerstein et al. 2010]; and cyclic internal processes in which the entire cell population can be synchronized [Spellman et al. 1998]. The next step is to determine the duration and sampling rates for the processes being studies.

2.3 Sampling rates

Sampling rates are closely related to the goals of the experiment. For example, if the experiment is about measuring a cyclic process, then the sampling should be uniform which means that the interval between consecutive time points should be the same. Besides, sampling should also cover multiple cycles in order to capture consistent changes. However, if the study is about development, then there is no simple and ideal sampling rates to follow. Two approaches are commonly used for developmental studies. The first approach depends on morphological markers of the embryo as a substitution for stages of transcriptional regulation [Gerstein et al. 2010]. The second approach is to change the sampling rate during the life cycle based on the expected rates of changes in gene expression.

After solving the problem of choosing an appropriate sampling rates, another problem is choosing the sampling density, which depends on the goal of the experiment. In general, under the constraint of budget, an important question is whether to invest in more replicates for each time point, or in more time points but fewer replicates for each time point. If the goal is to investigate the kinetic pattern of a biological process, it would be better to invest more in time points rather than replicates. A denser sampling can also help to control the noise in individual time points even without replicates. However, if the goal is to find differentially expressed genes across different time points, then it would be better to invest more in replicates.

In practice, choosing an appropriate sampling rate is difficult because of the limit of pre-knowledge of the biological process. One useful solution is to monitor the expression of a few genes over a long time period to try to find the pattern behind the biological process before choosing a sampling rate [Amit et al. 2009].

2.4 Synchronization of time-series gene expression experiments

Microarray-based and high-throughput sequencing-based experiments currently require a population of cells. Therefore, how to make sure that all the cells are in the same phase of the biological processes throughout all time points in the experiment, which is called synchronization, is an important issue.

An example of within-series synchronization is studying the cell cycle [Spellman et al. 1998]. Cell cycle study requires a synchronized population of cells at multiple time points during the cell cycle. The most commonly used synchronization method in these experiments is arresting cells at a specific

point in the cell cycle and then releasing them at the same time. This approach was initially used in budding yeast [Spellman et al. 1998].

However, although these arresting methods are effective in lower organisms with slow cell cycles, their application to mammalian cells, which usually have longer cell cycles, are ofter not effective because the cells will lose synchronization quickly. Even for yeast cells, such arrest methods did not lead to complete synchronization [Lu et al. 2004].

Various methods have been introduced for synchronizing cells in a cyclic experiment. Most of them rely on matching the phases for the first and second cycle for each gene [Lu et al. 2004]. This works well when at least the first cycle is fairly synchronized, however, they cannot be applied to mammalian cell cycle or to other responses in which the activity is not cyclic.

An alternative set of approaches is to synchronize cells *in silico*. One of these approaches relies on additional measurements to characterize the population of cells at each time point such as flow-cytometry-based analysis of DNA quantity. Using these measurements, a model for the population of cells at each time point is constructed and is then used to deconvolve the time-series expression data [Bar-Joseph et al. 2008].

Another type of synchronization is used when combining or comparing time-series experiments from multiple studies. In such cases, although each individual time-series dataset may be synchronized, response rates may differ between these datasets. Therefore, it is difficult to compare results between different time-series experiments. Several approaches have been developed to solve these problems. Most of them rely on the alignment of expression profiles between the experiments using a time-wrapping method [Kaminski and Bar-Joseph 2007] or use hidden Markov models for the alignment process [Lin et al. 2008].

2.5 RNA-seq time-series gene expression data

Although most time-series gene expression data sets are based on microarrays and microarray analysis methods are more mature, many time-series RNA-seq studies have been carried out over the past few years [Pauli et al. 2012]. There are several advantages of RNA-seq time-series gene expression data. First, some noise issues of microarrays such like background correction and cross-hybridization, are solved by RNA sequencing. Second, RNA sequencing studies make it easier to determine the expression of alternatively spliced genes, and provide opportunities for expression experiments of species that do not have an assembled genome. Third, sequencing-based methods are more replicable and lead to more accurate results compared with microarray-based methods [Marioni et al. 2008]. Although there are still problems with RNA-seq data, it is expected that over the next few years, most time-series expression data will be based on RNA-seq technology.

2.6 Clinical application

Time-series gene expression data are being increasingly used to monitor patient responses to injury and disease [Calvano et al. 2005], as well as to treatments and preventive measures [Baranzini et al. 2004] in clinical studies. Patient heterogeneity can make the analysis of absolute expression levels meaningless, and ethnicity can affect the responses to therapy. Therefore, time-series measurements that provide information about expression changes are especially beneficial. However, there are some unique challanges for such studies. For example, ethical considerations may preclude certain types of samples that would be most relevant to the scientific hypotheses. Choosing the correct sampling rate is also difficult as mentioned before. In many cases the transcriptional changes occur within days, but it may take years to see whether a patient has responded successfully to treatment or not, indicating that a longer duration of sampling would be more appropriate in such studies.

Task	Software	Description	Link
Identifying differentially expressed genes	Linear Models for Microarray Data (LIMMA)	Uses linear models to analyse gene expression and is part of the popular Bioconductor project	http://www.bioconductor.org/packages/ release/bioc/html/limma.html
	Significance Analysis of Microarrays (SAM)	Permutation-based analysis of gene expression	http://www-stat.stanford.edu/~tibs/SAM
	Extraction of Differential Gene Expression (EDGE)	Statistical analysis that specifically leverages the time structure in the expression data	http://www.genomine.org/edge
	Bayesian Estimation of Temporal Regulation (BETR)	Bayesian technique that exploits time-dependent structure in the expression data and is available with the MultiExperiment Viewer (MeV) application and Bioconductor	http://www.tm4.org/mev
Clustering	Short Time-series Expression Miner (STEM)	Maps genes to representative expression profiles with an emphasis on short time-series experiments; also implements k-means	http://www.sb.cs.cmu.edu/stem
	Graphical Query Language (GQL)	Hidden Markov model (HMM)-based clustering	http://ghmm.org/gql
	Cluster Analysis of Gene Expression Dynamics (CAGED)	Models gene expression using autoregressive equations	http://dcommon.bu.edu/xmlui/handle/2144/1290
	TimeClust	Implements hierarchical clustering, self-organizing maps, and two novel time-series clustering algorithms	http://aimed11.unipv.it/TimeClust
	Dynamic modelling and clustering (DynaMiteC)	Simultaneously clusters genes and fits groups of similar genes to impulse models	http://www.compbio.cs.huji.ac.il/DynaMiteC/Site/DynaMiteC.html
	Platform for Processing Expression of Short Time Series (PESTS)	Summarizes expression profiles with various features and can also identify significant genes	http://www.mailman.columbia.edu/ academic-departments/biostatistics/ research-service/software-development
Classification	GQL	Extensions of GQL enable it to classify clinical responses on the basis of gene expression	http://ghmm.org/gql
	Treatment-Response Alignment Models (TRAM)	Discriminative HMM-based classification	http://www.cs.cmu.edu/~thlin/tram
	MVQueries	Uses HMMs to model expression response as piecewise constant functions	http://bioinformatics.rutgers.edu/ Software/MVQueries
Dynamic regulatory networks	Inferelator	Ordinary differential equations are used to model transcriptional changes in terms of environmental and transcription factor influence	http://err.bio.nyu.edu/inferelator
	Network Component Analysis	Decomposes a dynamic gene expression matrix to learn transcription factor activities over time	http://www.seas.ucla.edu/~liaoj/ download.htm
	Dynamic Regulatory Events Miner (DREM)	HMM-based algorithm for identifying transcription factors that control divergence points in gene expression profiles	http://www.sb.cs.cmu.edu/drem
	Time-Series Network Identification (TSNI)	Constructs a local regulatory network of genes that are affected by an external perturbation	http://dibernardo.tigem.it/wiki/index.php/Time_Series_Network_ Identification_TSNI
Simulation	GeneNetWeaver	Generates realistic regulatory networks and dynamic gene expression data	http://gnw.sourceforge.net

Fig. 1. Software for the analysis of time-series gene expression data

The primary goal of many clinical studies is using classification to predict patient outcome. For example, in a study in which healthy volunteers were exposed to influenza, researchers used logistic regression to identify genes that can discriminate between pairs of phenotypic classes [Huang et al. 2011]. One problem about clinical studies is poor reproducibility, which inhibits the applicability of these studies to medical practice. Therefore, it is important to prevent overfitting the gene expression data. The methods we can use to prevent overfitting are: concentrating on small subsets of relevant genes based on prior knowledge [Baranzini et al. 2004]; using cross-validation strategies [Baranzini et al. 2004]; or validating results using independent patients and/or experimental methods [Calvano et al. 2005].

In the rest of this review, we will focus on computational methods that are designed for time-series experiments analysis. **Figure 1** lists several commonly used softwares for the analysis of time-series gene expression data. The software are classified according to different purposes of the analysis.

3. COMPUTATIONAL ANALYSIS OF TIME-SERIES GENE EXPRESSION DATA

3.1 Normalization

Normalization methods for time-series gene expression data are usually the same with static gene expression data because normalization is mainly focused on normalizing data in individual microarrays. However, there are some cases that normalization methods used for static expression data are not effective. One example is experiment aiming to measure RNA decay rates over time [Shalem et al. 2008]. Such experiments violate one primary assumption for most normalization methods: total quantity of mRNA is the same at different time points [Bolstad et al. 2003]. One optimal normalization method for such cases is to use spike controls. If spike controls are not available, some other normalization methods such like dChip [Li and Wong 2001] can be used because it does not rely on total RNA quantities. Methods like dChip rely on rank-invariant genes, which probably exist even after transcription shutdown.

3.2 Differentially expressed genes

After normalization, the question we then ask is how to identify differentially expressed genes. A heuristic solution that is commonly used is that a genes is differentially expressed if its expression value is above a chosen fold change in at least two consecutive points. However, the cutoff of these methods are arbitrarily chosen so that it may not be appropriate for all genes. To solve this problem, numerous methods have been developed, or extended, to identify differentially expressed genes in time-series data (**Figure 1**). Unlike the heuristic methods, these methods often rely on analyzing a more continuous version of the experiments data for each gene. Therefore, more time points are used to identify differentially expressed genes. The comparison between these methods and methods that are used for the analysis of static gene expression data such as t-tests indicates that in at least some cases, these methods can improve the identification of differentially expressed genes. Some of these methods require replicates at each time point but some do not.

3.3 Clustering

Although clustering methods for static gene expression data analysis such like hierarchical clustering and k-means clustering can be used to time-series gene expression data, there are some clustering methods specifically developed for time-series data. These methods include methods that use regression analysis to group genes on the basis of their trajectories [Ramoni et al. 2002]; methods based on graphical models like hidden Markov models (HMMs) to group genes on the basis of their transcriptional trends, regardless of the specific values [Schliep et al. 2003]; and methods that assign genes to

one of several previously defined temporal trajectories, therefore allowing users to determine significance levels for the different clusters [Ernst et al. 2005]. By including information at multiple time points, these methods are often an improvement on the static-based methods.

3.4 Classification

In studies of diseases, it has been observed that the dynamics of gene expression profiles may provide insights into the the serverity and the response of patients to treatments [Baranzini et al. 2004]. Many methods have been developed recently to analyze such data by classifying outcomes based on the dynamics of expression changes. It is shown that in many cases, by including information at multiple time points, these methods outperform methods that only using static data.

3.5 Causality

A key advantage of time-series data is that it can be used to infer causality without perturbing the system. Based on the dynamic change of expression profile across different time points, researchers can test several hypotheses regarding causal relationships between genes. Early work in this field used an alignment approach to match similar or opposite subsections of expression patterns that were temporally separated. These alignment methods were used to identify potential activators and repressors [Qian et al. 2001]. Several other methods use various types of regression analysis which can also be used to identify such causal relationships. In these methods, researchers try to model the expression profile of a specific gene based on the expression of another gene which is expressed earlier. Methods that use continuous representation are more appropriate for this type of analysis. Although most work in this field is focused on modelling organisms, dynamic Bayesian networks, which rely on the expression of a regulator at one time point to explain the expression of a target at the next, were successfully applied to identify causal candidates for the temporal changes in a human blood transcriptional network [Zhu et al. 2010]. However, due to the high dimensionality of data, false positives remain a major problem when carrying out such causality analysis. In addition, because many transcription factors are only post-transcriptionally regulated, such an analysis may miss key regulators. Therefore, a better approach is to integrate additional types of genomic data when carrying out such causal modelling.

4. CONCLUSIONS

Time-series gene expression data provides a wealth of information about the dynamics of gene expression, possible interactions between genes and the role that different genes play in a biological process. By integrating other static omics data sets, researchers can investigate the dynamic networks that are activated in cells from a global perspective. Given the importance of dynamic biological processes, the insights that are derived from current high-throughput dynamic data and the increased ability to study dynamic system, time-series data will play an even more important role in future studies.

5. AUTHORS CONTRIBUTION

Le and I collected information and read papers together. He is responsible for the presentation and I am responsible for writing the paper. Due to the 8 pages limitat of the paper, we decided to talk about specific methods and the details in the presentation but not in the paper. The paper is just a general review of time-series gene expression data. The overall contribution is 50% for each person.

REFERENCES

Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K Grenier, Weibo Li, Or Zuk, and others. 2009. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326, 5950 (2009), 257–263.

Ziv Bar-Joseph, Zahava Siegfried, Michael Brandeis, Benedikt Brors, Yong Lu, Roland Eils, Brian D Dynlacht, and Itamar Simon. 2008. Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proceedings of the National Academy of Sciences* 105, 3 (2008), 955–960.

Sergio E Baranzini, Parvin Mousavi, Jordi Rio, Stacy J Caillier, Althea Stillman, Pablo Villoslada, Matthew M Wyatt, Manuel Comabella, Larry D Greller, Roland Somogyi, and others. 2004. Transcription-based prediction of response to IFNβ using supervised computational methods. *PLoS biology* 3, 1 (2004), e2.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, and others. 2013. NCBI GEO: archive for functional genomics data sets?update. *Nucleic acids research* 41, D1 (2013), D991–D995.

Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 2 (2003), 185–193.

Steve E Calvano, Wenzhong Xiao, Daniel R Richards, Ramon M Felciano, Henry V Baker, Raymond J Cho, Richard O Chen, Bernard H Brownstein, J Perren Cobb, S Kevin Tschoeke, and others. 2005. A network-based analysis of systemic inflammation in humans. *Nature* 437, 7061 (2005), 1032–1037.

Jason Ernst, Gerard J Nau, and Ziv Bar-Joseph. 2005. Clustering short time series gene expression data. *Bioinformatics* 21, suppl 1 (2005), i159–i168.

Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* 11, 12 (2000), 4241–4257.

Mark B Gerstein, Zhi John Lu, Eric L Van Nostrand, Chao Cheng, Bradley I Arshinoff, Tao Liu, Kevin Y Yip, Rebecca Robilotto, Andreas Rechtsteiner, Kohta Ikegami, and others. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330, 6012 (2010), 1775–1787.

Yongsheng Huang, Aimee K Zaas, Arvind Rao, Nicolas Dobigeon, Peter J Woolf, Timothy Veldman, N Christine Øien, Micah T McClain, Jay B Varkey, Bradley Nicholson, and others. 2011. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS genetics* 7, 8 (2011), e1002234.

Naftali Kaminski and Ziv Bar-Joseph. 2007. A patient-gene model for temporal expression profiles in clinical studies. *Journal of Computational Biology* 14, 3 (2007), 324–338.

Cheng Li and W Hung Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2, 8 (2001), 1–11.

Tien-ho Lin, Naftali Kaminski, and Ziv Bar-Joseph. 2008. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics* 24, 13 (2008), i147–i155.

Xin Lu, Wen Zhang, Zhaohui S Qin, Kurt E Kwast, and Jun S Liu. 2004. Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic acids research* 32, 2 (2004), 447–455.

John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 18, 9 (2008), 1509–1517.

Andrea Pauli, Eivind Valen, Michael F Lin, Manuel Garber, Nadine L Vastenhouw, Joshua Z Levin, Lin Fan, Albin Sandelin, John L Rinn, Aviv Regev, and others. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research* 22, 3 (2012), 577–591.

Jiang Qian, Marisa Dolled-Filhart, Jimmy Lin, Haiyuan Yu, and Mark Gerstein. 2001. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of molecular biology* 314, 5 (2001), 1053–1066.

Marco F Ramoni, Paola Sebastiani, and Isaac S Kohane. 2002. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences* 99, 14 (2002), 9121–9126.

Alexander Schliep, Alexander Schönhuth, and Christine Steinhoff. 2003. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics* 19, suppl 1 (2003), i255–i263.

Ophir Shalem, Orna Dahan, Michal Levo, Maria Rodriguez Martinez, Itay Furman, Eran Segal, and Yitzhak Pilpel. 2008. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular systems biology* 4, 1 (2008).

Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. 1998. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* 9, 12 (1998), 3273–3297.

Jun Zhu, Yanqing Chen, Amy S Leonardson, Kai Wang, John R Lamb, Valur Emilsson, and Eric E Schadt. 2010. Characterizing dynamic changes in the human blood transcriptional network. *PLoS computational biology* 6, 2 (2010), e1000671.