Time-Frequency Feature Detection for Time-course Microarray Data

Jiawu Feng*
Courant Institute Bioinformatics Group,
New York University
Jiawu@cs.nyu.edu

Paolo Emilio Barbano*
Courant Institute Bioinformatics Group, New York University &
Dept of Math, Yale University
peb22@pantheon.yale.edu

Bud Mishra
Courant Institute Bioinformatics Group, New York University &
Cold Spring Harbor Laboratory
mishra@nyu.edu

ABSTRACT

Gene clustering based on microarray data provides useful functional information to the working biologists. Many current gene-clustering algorithms rely on Euclidean-based distance metrics and fail to capture the time-dependent features of the data, usually corrupted by high levels of experimental noise. Here we propose an algorithm capable of dealing with the noise through a time-frequency approach and related measure of correlation between time-course expressions of different genes (trajectories). The approach makes use of fast multi-resolution feature classification algorithms and allows for the desired functional characteristics (such as phase delay, activation/repression etc.) to be enhanced and detected.

We have applied our algorithm to time-course microarray data of ***Drosophila melanogaster*** (Arbeitman *et al.*, Science, Sep 27, 2002, page 2270-2275). We examined various relations among homeodomain genes (referred to as group H) and regulators of homeodomain genes (group RH) as follows: After normalization, the trajectories were projected on to CosBell wavelet basis. The four genes in group RH form two clusters: three of them stayed close to each other, and the last one, CG8651 (trithorax), was singled out. The group H genes, forming four clusters, showed functional features that are more similar to trithorax than the other three. We further analyzed ten homeodomain genes that have good correlations with trithorax in the wavelet basis. Literature search showed that there are five genes thought to be in the downstream pathway of trithorax. Although only two of these five genes were in the dataset available to the algorithm, it was able to identify both of these. Our study suggests that time-frequency analysis provides a powerful tool for discovering the underlying regulatory networks when applied to time-course microarray data.

Categories and Subject Descriptors

[**Bioinformatics**]: Clustering of very large dimensional data such as those from microarrays and proteomic experimental platforms.

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Time Frequency Analysis, Local Distance, Gene Networks, Functional Genomics

1. INTRODUCTION

One of the fundamental problems of cell biology is to understand how genes behave individually and how the features of different genes interact to carry out complex biological functions. Traditionally, biologists investigate the functions of genes by focusing on handful of genes each time. Recent advances in the microarray technology have made it possible to simultaneously measure the mRNA expression level of thousands of genes. Given such large amount of data, computational and mathematical techniques became essential for the correct interpretation of such large data sets. A variety of machine learning methods, both supervised and unsupervised, has been applied to microarray data. Since the underlining structure of the gene network is largely unknown and building labeled data sets for supervised learning is difficult, unsupervised methods are more popular in the research community. Current unsupervised clustering methods includes hierarchical clustering, self-organizing maps, relevance networks, principal components analysis, nearest neighbors, support vector machines, etc. All these clustering methods are based on certain types of measures of distances (metrics) between genes, such as Euclidean distance, Pearson correlation coefficients, and mutual information. For a detailed treatment of relative advantages and disadvantages of these techniques, please refer to [1]. The metrics developed through the methodologies mentioned above are not ideal, as they obscure many interesting biological features of the data: Euclidean distance brings up complicated normalization problems, and it is not robust to noise; Pearson correlation coefficients rely on normal densities of the measurements and

SAC '04, March14-17, 2000, Nicosia, Cyprus.

The work reported in this paper was supported by grants from DARPA's BioCOMP project (Title: "Algorithmic Tools and Computational Frameworks for Cell Informatics") and AFRL contract (contract #: F30602-01-2-0556). Additional support was provided by NSF's Qubic program, HHMI biomedical support research grant, the US department of Energy, the US air force, National Institutes of Health and New York State Office of Science, Technology & Academic Research

*These authors contribute equally to this work.

linear models of interactions; mutual information depends on the number of 'bins' used, while such 'bins' can be very difficult to identify correctly [1].

In this work, we propose a different approach to the problem of establishing a meaningful notion of distance for time-coursed gene expression data. The method requires the number of samples to be relatively large (at least a dozen, depending on the data-set).

We consider time-series data (trajectories) as mathematical functions within a larger system, and identify the relationship between these functions by means of time-frequency analysis and "network-correlations". We have applied this method to the time-course microarray data of Drosophila's development [2] and discuss our results in a later section. Our results suggest that this kind of analysis can be a powerful tool for measuring the correlations of gene expressions within the context of the gene network they operate in.

2. COMPUTATIONAL FRAMEWORK

The basic assumption underlying our technique, is that genes derive their functionality from the role assigned to them in a network of interacting genes. In order to produce an efficient algorithm to understand these functions, we have to effectively *translate* their biological function into mathematical relations and identify the candidate genes that facilitate the transclation process.

In most cases the group of candidate genes will be already known from the biological context. Other methods can be considered as well. The next section offers a strategy to deal with this problem.

2.1. Adaptive Basis Selection

One possible way to identify an initial set of genes for functional analysis is as follows: Focus on a small specific set of Time-Frequency features (such as highly localized oscillatory behavior etc.) and extract those genes exhibiting the required characteristics by means of Multi-resolution classifiers. One such classifier we explore here is a variant of the so-called Local Discriminant Basis [7].

The primary objects of consideration are finite sets of functions of the form $F = \{f(t), 0 \le t \le T\}$, along with their approximate representations in terms of *M*-dimensional vectors in a Euclidean space:

$$\widetilde{F} = \left\{ \widetilde{f}_i(t) = \sum_j < f_i, g_j > g_j(t) = \sum_j c_{ij} g_j(t), 1 \le j \le M \right\} \quad (1)$$

Such a vector representation is referred to as the "projection" of the time series F. The choice of the subset $\widetilde{B} = \{g_j(t), 1 \le j \le M\}$ of an orthonormal basis for $L^2[0,T]$ is of fundamental importance in order to capture the desired features of the data. More specifically, an appropriate choice of such $\widetilde{B}$ will suffice for the Euclidean distance between the projections of two sets of time series functions to determine if such functions, in fact, describe similar behavior of the system.

In the ideal case, one has many such time series functions and a natural choice of B and $\widetilde{B}$ can be made so that, once the functions have been projected onto a finite dimensional Euclidean space, the most typical as well as robust behaviors of the system can be determined by those functions whose projections all lie within, say, n "small" Euclidean spheres, $\{B(x_1, \varepsilon), \ldots, B(x_n, \varepsilon)\}$, with the property that

$$x \in B(x_j, \varepsilon), y \in B(x_k, \varepsilon) \Rightarrow \|x - y\| > 3\varepsilon \quad (2)$$

i.e., the sets of time series functions giving rise to unique clusters.

Thus, in order to apply this method effectively to analyze biological trajectories, it only requires that suitable orthonormal bases have been selected for a biological process under examination. It is further desirable that the analysis can be carried out with a feasibly small value of *M* (say *M*=2 representing the Euclidean plane) and suitable ε. Our algorithm consists of a wavelet-based algorithm to devise an appropriate orthonormal bases and subsequently, compute the projections. The examples we considered demonstrate that the method is applicable for a vast number of biological processes and requires only projections on to the plane (*M*=2).

The next issue to be considered is to identify the role that the selected genes play inside the network they are imbedded in.

2.2. Functional Correlation Sets

Next, we introduce a notion of "network-correlation" of a pair of genes (g_j, g_k), belonging to a gene network $N = \{g_i\}_{1<i<n}$. We proceed as follows: the time trajectories of the pair (g_j, g_k) are re-sampled and filtered to obtain two slightly smoother, yet completely faithful representations of the original pair. The re-sampled genes are then normalized in the square norm. We denote the resulting new pair with $(\widetilde{g}_j, \widetilde{g}_k)$. The functional correlation matrix $\underline{C}_{jk}$ with respect to N is defined as:

$$\begin{aligned} \alpha_{j,l} &= \alpha_{l,j} = \left\| \widetilde{g}_j \bullet \widetilde{g}_j - \widetilde{g}_l \bullet \widetilde{g}_l \right\| \\ C_{jk} &= \{\alpha_{j,l}, \alpha_{k,l}\}_{l \in N} \end{aligned} \quad (1)$$

Where $\bullet$ denotes the cyclic correlation of the vectors and the norm is taken in the Euclidean sense. In doing so we have associated a $n \times 2$ matrix to the pair. This new set contains information to understand how the two genes are acting on the network with respect to each other. There are two essential aspects to this simple computational procedure:

- High robustness with respect to additive as well as phase noise (i.e. time-shifts/dilations of the signals with respect to each other). This allows for experimental errors to be absorbed very well.
- High robustness with respect to localized frequency perturbations. This feature may be crucial to deal with "burst errors" (due for example to short-time systematic perturbations) in some of the trajectories.

The next step in the algorithm is to identify geometric features of these Functional Correlation Sets (FCS), viewed as point in the

Euclidean plane, and associate the corresponding biological function to the genes that generate them.

3. BIOLOGICAL ANALYSIS OF FUNCTIONAL CORRELATION SETS (FCS)

We first selected two groups of genes from the data in [2]: homeodomain (GroupH) genes and their regulators (GroupRH). GroupRH consists of four genes E(z), ash2, esc and trx. E(z) and esc belong to a group of proteins referred as Polycomb Group (PcG). These proteins bind to a DNA fragment of several hundred base pairs, which is called Polycomb response elements (PREs). PcG genes are responsible for maintaining repression state of homeodomain genes during *Drosophila* early development. Interestingly enough, Trx and related proteins (trx-group, or trx-G) also bind to PREs, but their effect is the opposite of PcG: they maintain the derepression state (active state) of homeodomain genes expression. Whether the target gene is repressed or derepressed depends on the preset of earlier regulators, the jobs for PcG and Trx are just to keep the memory of previous states [3]. It has also been reported that E(z) is required for binding of Trx and other proteins to specific chromosomal sites where they may interact with other chromatin factors to alter target gene transcription [4]. Ash2 belongs to trx-G. It is also reported that in yeast, homologs of *Drosophila* Ash2 and Trx form a protein complex called SET1 with the function of reforming chromatin structure [7].

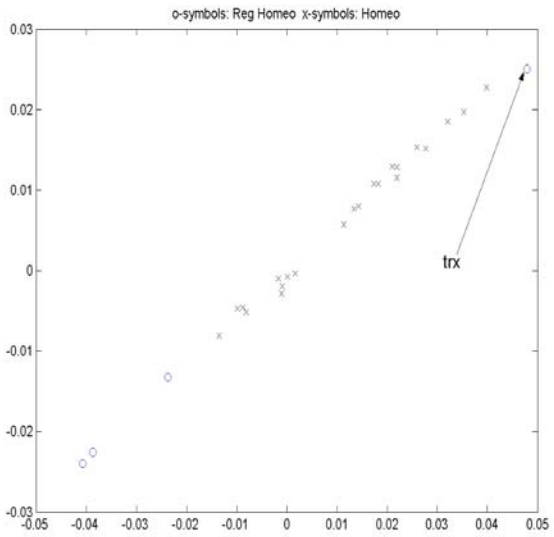
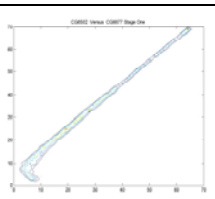
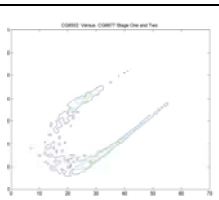
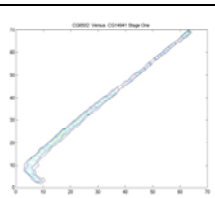
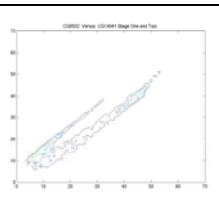
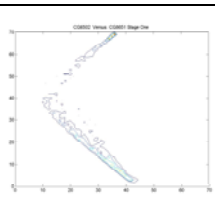
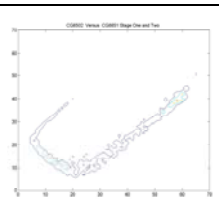
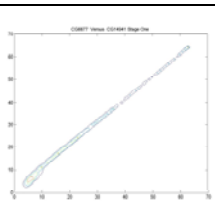
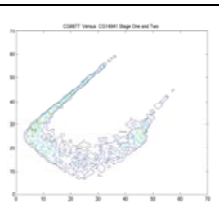
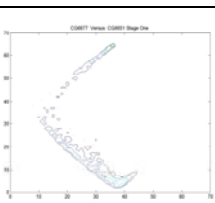
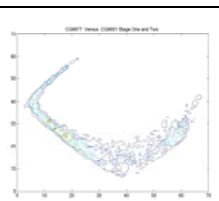
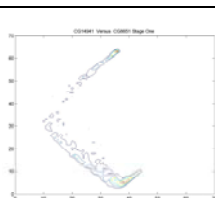
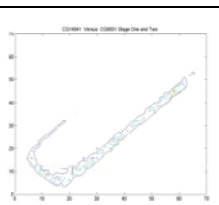


Figure 1. Plot of the first two most important Time-Frequency components of the GroupRH(Circles) and GroupH (Cross) genes. The point corresponding to trx appears very distant from the other three in its group.

We proceeded as follows. First, we isolated Time-frequency features of the GroupH and GroupRH by means of cosine-bell (CosBell) wavelet-packets and performed their clustering analysis. The result clearly indicated the drastic difference between trx and the other genes in GroupRH. The four genes in group RH form two clusters, three of them stayed close to each other, whereas the last one, CG8651 (also called trithorax, or trx), is singled out. It is interesting to observe that the GroupH genes, while forming four clusters, were displaying Time-Frequency features similar to the ones of trx. Only two of the GroupH genes have been suggested in the literature to be in the downstream of trx; these two genes are AntP and adbA, which were found to be closely related to trx in the time-frequency analysis.

Table 1. Summary of shapes (in contour map) in the correlation analysis. Abbreviations: ES (Early Stage); ELS (Early + Lava Stage); ISC (Is Shape Changed?)

Pairs	ES	ELS	ISC
E(z)-ash2			No
E(z)-esc			No
E(z)-trx			Yes
Ash2-esc			No
Ash2-trx			Yes
Esc-trx			No

The next step consisted in creating the FCS (Functional Correlation Sets) for the GroupRH genes and detecting their functional relations. We used a simple graphical analysis by plotting our N by 2 matrices onto a contour map, where we can compare the density distributions of the other genes in the network with respect to the particular pair of genes. We summarized the results of our correlation analysis in Table 1. For the full-set result and more detailed explanation, please refer to the on-line supplementary materials at:
http://www.cs.nyu.edu/cs/faculty/mishra/NOTES/mynotes.html.

4. CONCLUSIONS

By combining the results of Time-Frequency analysis (Figure 1) and FCS (Functional Correlation Sets) analysis (Table 1) with biological knowledge, we conclude the following:

1. The geometric features of the FCSs (Functional Correlation Sets) indicated that trx has an antagonistic relation with E(z), esc, and ash2.

2. The expression levels of E(z), esc, ash2 are very consistent throughout the 'early + lava' stage, which may suggest that they form a stable protein complex. Such a complex is confirmed by several studies ([4], [3], [7]). It is not surprising that E(z) and esc have similar shapes, since they cooperate as a repression mechanism. However, the behavior of ash2 is somehow mysterious, since it is reported to belong to the Trithorax-Group and has a function that is opposite to those of E(z) and esc [7].

3. The contour shapes of two pairs containing trx changed between early and 'early + lava'. Which suggests that the behavior of trx is different from the other genes. This is consistent with our observations from Time-Frequency analysis.

4. Considering point 2 and 3, we speculate that although ash2 is supposed to be a de-repressor, it might not function by itself. The scenario could be that ash2 was a static component of the protein complex and might cooperate with a dynamic component (such as trx) to de-repress genes transcription.

5. Although PcG and trx-G have opposite effects on homeodomain gene expression, their logical status might not be equivalent: PcG appears more like a static, "default" configuration and trx-G appears more like a dynamic, "alternative" configuration.

5. DISCUSSIONS

Understanding the complex genetic networks at the cell biology level is a crucial task for the biologists and is of enormous biomedical value as well. Due to the limitations of current biotechnological systems, such a mission cannot be accomplished in one single step. A more viable approach is to gather many pieces of information about a network through high-throughput experiments, and then computationally put things together later on. Microarray analysis of gene expression profiles provides much such useful information by direct comparison of "normal" state and "alternative state" of the target organism and by more advanced studies such as gene clustering.

Nowadays, the popular gene clustering algorithms often give large groups of clusters that often contain more than a hundred genes. Such large clusters make biological validation a prohibitive task. Here we emphasize more on a specific group of genes, hence can give results that are provable by established biological experiments such as RNAi, gene knockouts/knock-ins or yeast two-hybrid experiments. In addition, mathematically, we can "deconvolve" the time-course microarray data to provide very useful information that non-time-course data lack. An explanation for this added informativeness is that time-course data clearly reflect the internal natural constraints imposed on biology, whereas scattered sampling of genes expression obscure such information. Furthermore, classical statistical analysis grounged on the assumptions of "laws of large numbers" views gene expression as a collection of a large number of independent random events (patently false in biology) and thus "looses the context" in the sense that the expression of entire set of genes in an individual organism is a system. Our correlation analysis, on the other hand, takes the existence of such a system into consideration in order to assign a functional meaning to a gene. For these reasons, we believe that large-scale multi-resolution geometric analysis of time-course data will occupy a central position in systems biology.

6.REFERENCES

[1] Butte, A. The use and analysis of microarray data. Nature reviews drug discovery 1 (2002), 951-959.

[2] Arbeitman, M. N., Furlong, E. EM., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., White, K. P. Gene Expression during the Life of Drosophila Melanogaster. Science 297 (2002), 2270-2275.

[3] Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites. Cell 111 (2002), 185–196.

[4] Breen, T.R. Mutant Alleles of the Drosophila trithorax Gene Produce Common and Unusual Homeotic and Other Developmental Phenotypes. Genetics 152 (1999), 319–344.

[5] Beltran, S., Blanco, E., Serras, F., Pérez-Villamil, B., Guigó, R., Artavanis-Tsakonas, S., and Corominas, M. Transcriptional network controlled by the trithorax-group gene ash2 in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA 100 (2003) **,** 3293-3298.

[6] Nagy, P. L., Griesenbeck, J., Kornberg, R. D. and Cleary M. L. A trithorax-group complex purified from Saccharomyces cerevisiae is required for methylation of histone H3. Proc. Natl. Acad. Sci. USA 9 (2002), 90-94.

[7] Coifman, R. R. and Saito N. Local Discriminant Bases and their Applications. Journal of Mathematical Imaging and Vision 5 (1995), 337-358.

[8] Breen, T. R., and Harte, P. J. Molecular characterization of the trithorax gene, a positive regulator of homeotic gene expression in Drosophila. Mech. Dev. 35 (1991), 113-127.