SELECTING MAXIMALLY INFORMATIVE GENES TO ENABLE TEMPORAL EXPRESSION PROFILING ANALYSIS

I.P. Androulakis[1,2,*], J. Vitolo[2] and C. Roth[1,2]
[1]*Biomedical Engineering Department*
[2]*Chemical & Biochemical Engineering Department*
Rutgers, The State University of New Jersey, Piscataway, NJ 088554

Abstract

Use of microarrays to analyze drug responses has mainly been restricted to comparing treated versus untreated samples at a few time points. In order to decipher the complex expression interactions and their biological implications, we need to account for the temporal evolution of expression profiling over a number of time points. This paper analyzes data obtained from an extended microarray time series (rat liver) for the in vivo responses to a single dose of methylprednisolone (Almon, DuBois et al. 2003). We propose a framework that identifies, in an unsupervised, exact and rigorous manner, distinct expression patterns and furthermore, assigns a critical subset of informative genes to each of these expression motifs. The biological relevance of the identified groups of informative genes is evaluated within the context of known biological mechanisms of corticosteroids and potential mechanisms suggested by analysis of gene promoter sequences.

Keywords

Expression profiling, regulatory architecture, motif discovery.

Introduction

It has been hypothesized that expression profiling using gene arrays can be used to distinguish temporal patterns of changes in gene expression in response to a drug in vivo, and that these patterns can be used to identify groups of genes regulated by common mechanisms. With the high throughput sequencing of the complete genomes of a variety of species almost complete, experimental strategies combined with enhanced advances in modeling and computing have allowed biologists to accelerate the pace of understanding gene expression and transcriptional regulation in a systematic manner. Using DNA microarrays patterns of similar expression profiles have been linked to shared regulatory mechanisms (Wei, Liu et al. 2004). The next challenge thus becomes to elucidate the function of these genes and to discover how they interact to perform specific biological processes, especially for the large fraction of genes in the genome whose functions are currently unknown (Stuart, Segal et al. 2003). A common approach is to exploit relationships among co-expressed genes as these may provide strong evidence for the involvement of new genes in known biological functions. Of paramount importance in elucidating the functionality of genes and their overall contribution in biological functions is the fundamental understanding of the intricate and precise regulatory process that provides living cells with their remarkable properties. Therefore, charting gene regulatory networks is a major focus of interest in modern biology; that is assessing the information transfer from regulatory genes to structural genes whose products account for the phenotypic response of the gene. Therefore, the compendia of available gene expression experiments and the totality of possible co-expressed genes, are being augmented by considering the entirety of the regulatory networks affecting the transcription process (Wasserman and Sandelin 2004; Dohr, Klingenhoff et al. 2005).

* Author to whom correspondence should be addressed: Ioannis (Yannis) P. Androulakis, Biomedical Engineering Department, Rutgers University, 617 Bowser Road, Piscataway, NJ 08854 (E-mail: yannis@rci.rutgers.edu)

For such an analysis two critical questions need to be addressed:

i. How to establish the exact link between genes that exhibit strong correlation in terms of their expression patterns and their underlying regulatory architecture (Qian, Lin et al. 2003) .
ii. How to establish the potentially complete patterns of co-expression and their non-intuitive temporal relations (Qian, Dolled-Filhart et al. 2001)

The subject of this paper is to explode an alternative approach towards the analysis of temporal expression data in an attempt to better characterize the nature of expression patterns. We also present some preliminary analysis in our attempt to establish relations among gene in terms of both their transcriptional profiles as well as their underlying regulatory architecture. In this work we focus on the work of (Almon, DuBois et al. 2003), analyzing the corticosteroid effects on rat liver.

Temporal Patterns of Drug response

Corticosteroids are a common group of drugs used to treat a variety of pathologies requiring anti-inflammatory intervention. A prerequisite to understanding the complexities of drug treatment is a broader identification of both genes affected by steroid treatment and the temporal patterns of transcriptional changes that occur (Almon, DuBois et al. 2003). Thus two issues become important: gene selection and identification of their associated major expression motifs.

(Almon, DuBois et al. 2003) 43 animals received a single 50-mg/Kg dose of methylprednisolone (MPL) sodium succinate (a corticosteroid). Two to three rats were sacrificed at the following time points after MPL administration: 0.25, 0.5, 0.75, 1, 2, 4, 5, 5.5, 6, 7, 8, 12, 18, 30, 48 and 72 h. Four animals were used as controls and were sacrifices untreated. RNA extracted from liver tissue at each time point was analyzed. The temporal expression patterns of 8799 genes at all time points were collected. Particular emphasis was placed on 197 corticosteroid responsive probe sets representing 143 different genes. Using a combination of supervised methods, self-organizing maps and k-means clustering, the expression patterns were classified into six patterns. In a subsequent work (Jin, Almon et al. 2003) pharmacokinetic models were derived for each of the six clusters to describe possible inhibitory and stimulatory mechanisms defining the transcription process.

A critical component in this, and other similar studies, is how to establish the relationship between the available transcriptional profiles and how to extract in a systematic and unsupervised manner significant expression motifs and the key gene subsets associated with each pattern. In this work we propose a very efficient algorithm for extracting significant patterns of expression (motifs), evaluate the groups proposed by (Almon, DuBois et al. 2003) and begin the assessment of the the relationships between gene characterized by similar expression motifs and their underlying regulatory architectures.

Clustering Temporal Gene Expression Data

Clustering of time series data, of which a subset is the transcriptional data from large-scale microarray experiments, is a very active area of research and a variety of problems have been discussed in the open literature. The fact that this problem persists, in particular as it related to genomic data, is just an indication of the many complexities both computational and interpretational. Among the leading candidates for clustering expression profiles are distance-based methods, with k-means clustering being one of the leading candidates. However, it has been argued recently that distance based methods generate local solutions that are not necessarily meaningful (Lin and Keogh 2004). Furthermore, identifying a priori the number of necessary clusters remains, in general, an open problem. However, significant successes have been identified in the open literature.

The purpose of this study is to explore alternative ways that attempt to characterize, in an unsupervised manner, the raw expression data in an attempt to identify leading “motifs” within the expression data. Our main motivation is to define identifiers that uniquely characterize each transcriptional profile. Our goal is to identify those transcripts that share significant components of their expression patterns.

Finding motifs in time series and proximity preserving hashing

The goal of our approach is to concurrently achieve a characterization of the transcriptional data as well as a significant dimensionality reduction in order to assess the qualitative characteristics of the expression data. In order to do so, we explore the idea proposed by (Lin, Keogh et al. 2002). The algorithm transforms the time series data into a sequence of symbols, which are subsequently hashed to unique (motif-dependent) identifiers. The hashing function explores the concept of proximity preserving hashing (Chin 1994), that is similar structures hashing to similar values. The steps of the algorithm are

1. Let $\left\{E_{i,t}\right\}, i = 1,\ldots,N; t = 1,\ldots,T$ be the expression data for gene “i” at time t
2. Transform each series to N(0,1) as $\tilde{E}_{i,t} = \dfrac{E_{i,t} - \mu_i}{\sigma_i}$
3. Partition the time horizon (T) into N_w windows
4. Generate a piecewise aggregate approximation of the series as follows: $\overline{E}_{i,t'} = \dfrac{N_w}{T} \sum_{j=\frac{T}{N_w}(t'-1)+1}^{\frac{T}{N_w}t'} \tilde{E}_{i,j}$
5. Discretize the piecewise linear approximation based on break points according to a normal distribution.

Break points are defined as the equi-probability partitions of the N(0,1) distribution. The normalized time series is transformed to a symbolic representation as follows: $E^{S}_{i,t'}, t' = 1,\ldots,\frac{T}{N_w}$. For example, a symbolic representation using α=3 symbols {1,2,3} would be such that $E_{i,t'} \le -0.43 \rightarrow E^{S}_{i,t'} = 1, E_{i,t'} > 0.43 \rightarrow E^{S}_{i,t'} = 3, E^{S}_{i,t'} = 2$ otherwise.

6. Assign a unique hash value (motif identifier) to each sequence using the following (Lin, Keogh et al. 2002):

   $h(T, N_w, \alpha) = 1 + \sum_{i=1}^{N_w} \left(ord(E^{S}_{i,t'}) - 1 \right) * \alpha^{i-1}$, α is the size of the alphabet used in the symbolic representation (Step 5). This hash functions assigns to each symbolic representation an integer in the interval [1, w^{α}]
7. The hash values (motifs) can now be sorted and similar motif values correspond to similar transcriptional profiles.

The overall process is illustrated in Figure 1 using as example one of the transcriptional profiles of our dataset.

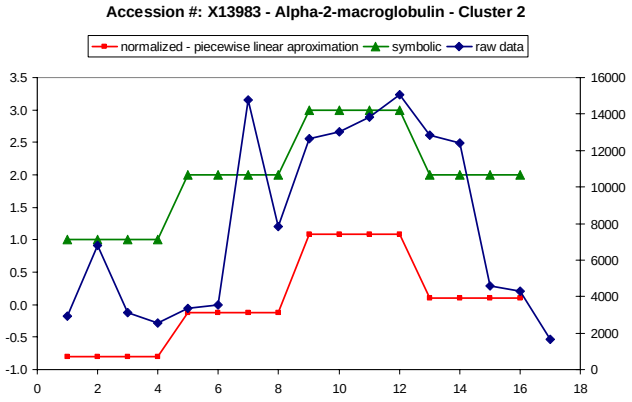

Figure 1: The representation scheme

Analysis of the (Almon, DuBois et al. 2003) Data

The dataset contains 8,799 probes, 17 times points with multiple repeats for each point. In total there are 47 measured values for each probe. As a first approximation, the repeats are average at each time point and the temporal transcriptional data contain 17 values per probe. The aforementioned analysis is extremely fast (about 1.5s). The following is a short summary of the computational results:

1. The original work of (Almon, DuBois et al. 2003) through a combination of supervised and k-means algorithms hypothesized the existence of 6 clusters. Our analysis identifies a potentially significant number of expected motifs. This is a further proof of the necessity of approaches such as the one proposed in this paper. If the actual number of expected motifs were unknown, distance-based algorithms that assume a knowledge of the potential number of clusters would essentially produce results that meet the user criteria.
2. Figure 2 depicts the "homogeneity" of each motif in terms of their corresponding values. Based on this we argue that cluster 1 forms a well defined neighborhood, and clusters 4 and 6 are very close in terms of their corresponding motifs and the actual transcriptional profiles are very hard to distinguish
3. We focused out analysis on the six clusters proposed by (Almon, DuBois et al. 2003). Based on our analysis, we confirm the existence of significant patterns (including clusters 1, 2, 3 and 5) with the most uniform distribution of motif values within the classes and also providing the strongest signatures of symbol distribution as shown in Figure 3. Clusters 4 and 6 are harder to distinguish

Conclusions

In this paper we have illustrated the application of a novel way for representing the information content of transcriptional profiles. We attempted to develop emerging motifs of the expression profiles and by analyzing those identify probes with persistent and overpopulated expression patterns. These in turn can be used to postulate tentative significant expression motifs characteristic of the transcriptional profiles. We are currently in the process of developing the associations between emerging stable expression profiles, participating genes and their possible common elements of regulatory elements. Based on the hypothesis that co-expressed genes are co-regulated based on the presence of common regulatory elements (or combinations thereof), we attempted to identify and characterize the gene promoter regions from clusters of co-expressed genes. For each of the clusters identified in (Almon, DuBois et al. 2003) we attempted to identify the promoter region; this was not possible in all cases. For those genes whose promoters could be identified, we analyzed the promoters for the presence of regulatory elements using the ModelInspector utility of the Genomatix software suite (http://www.genomatix.de). The 374 DNA-protein binding matrices (models) analyzed

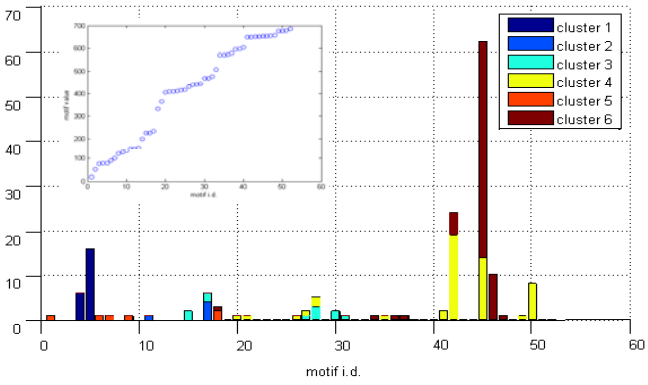

Figure 2: Motifs and corresponding clusters

were collapsed onto 142 families (functionally related binding sites). We are in the process of analyzing the possible correlations between genes belonging to similar motif families and the corresponding regulatory architectures. Preliminary results indicate possible relations between genes of similar motif and their regulatory network but the relations need to be further explored

References

Almon, R. R., D. C. DuBois, et al. (2003). "Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver." Funct Integr Genomics **3**(4): 171-9.

Chin, A. (1994). "Locality-Preserving Hash Functions for General-Purpose Parallel Computation." Algorithmica **12**(2-3): 170-181.

Dohr, S., A. Klingenhoff, et al. (2005). "Linking disease-associated genes to regulatory networks via promoter organization." Nucleic Acids Res **33**(3): 864-72.

Jin, J. Y., R. R. Almon, et al. (2003). "Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays." J Pharmacol Exp Ther **307**(1): 93-109.

Lin, J. and E. Keogh (2004). "Finding or not finding rules in time series." Applications of Artificial Intelligence in Finance and Economics **19**: 175-201.

Lin, J., E. Keogh, et al. (2002). Finding Motifs in Time Series. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada.

Qian, J., M. Dolled-Filhart, et al. (2001). "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions." J Mol Biol **314**(5): 1053-66.

Qian, J., J. Lin, et al. (2003). "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data." Bioinformatics **19**(15): 1917-26.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-55.

Wasserman, W. W. and A. Sandelin (2004). "Applied bioinformatics for the identification of regulatory elements." Nat Rev Genet **5**(4): 276-87.

Wei, G. H., D. P. Liu, et al. (2004). "Charting gene regulatory networks: strategies, challenges and perspectives." Biochem J **381**(Pt 1): 1-12.

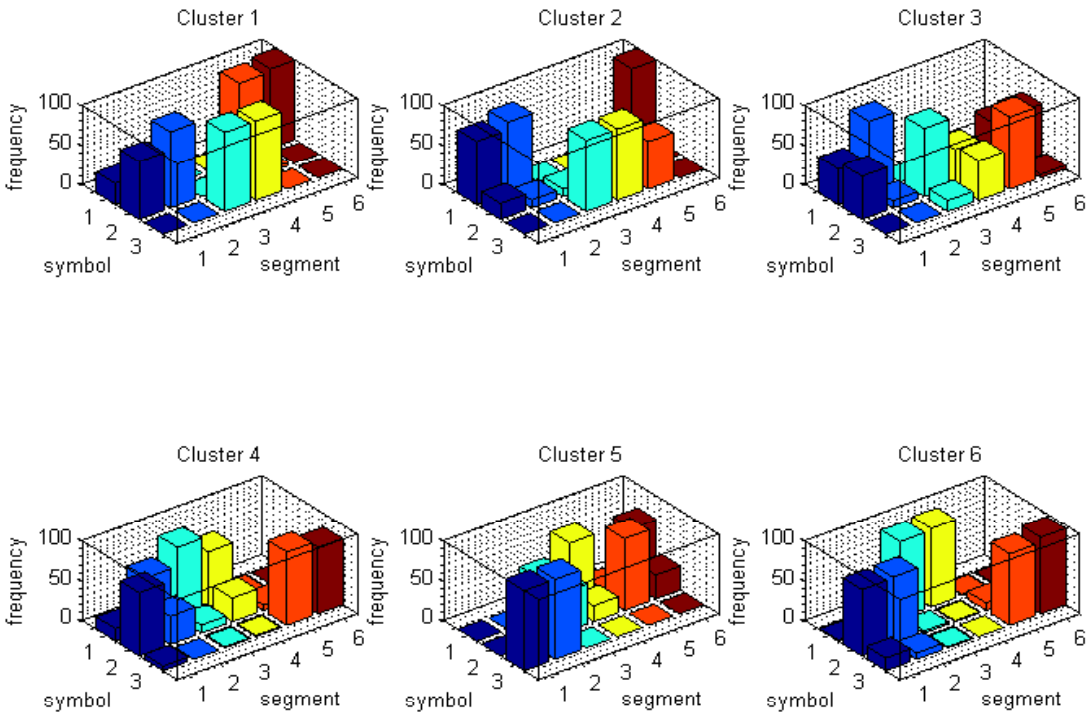

Figure 3: Distribution of motif values within the 6 clusters